44816

National Library of Canada
Canadian Theses Division
Ottawa, Canada
K1A 0N4

Bibliothèque nationale du Canada
Division des thèses canadiennes

PERMISSION TO MICROFILM — AUTORISATION DE MICROFILMER

• Please print or type — Écrire en lettres moulées ou dactylographier

Full Name of Author — Nom complet de l'auteur

Donald P. Sheridan

Date of Birth — Date de naissance	Country of Birth — Lieu de naissance
3 Aug 47	New Zealand

Permanent Address — Résidence fixe

Box 70
Hanna, Tan Park
Alberta

Title of Thesis — Titre de la thèse

The Effects of Feedback on Test Achievement in OAI

University — Université

Alberta

Degree for which thesis was presented — Grade pour lequel cette thèse fut présentée

PhD

Year this degree conferred — Année d'obtention de ce grade	Name of Supervisor — Nom du directeur de thèse
1980	E.W. Romaniuk

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to microfilm this thesis and to lend or sell copies of the film.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de microfilmer cette thèse et de prêter ou de vendre des exemplaires du film.

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

Date	Signature
April 21, 1980	Donald P. Sheridan

NL-91 (4/77)

National Library of Canada
Collections Development Branch

Canadian Theses on
Microfiche Service

Bibliothèque nationale du Canada
Direction du développement des collections

Service des thèses canadiennes
sur microfiche

NOTICE

The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

THIS DISSERTATION
HAS BEEN MICROFILMED
EXACTLY AS RECEIVED

AVIS

La qualité de cette microfiche dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de mauvaise qualité.

Les documents qui font déjà l'objet d'un droit d'auteur (articles de revue, examens publiés, etc.) ne sont pas microfilmés.

La reproduction, même partielle, de ce microfilm est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30. Veuillez prendre connaissance des formules d'autorisation qui accompagnent cette thèse.

LA THÈSE A ÉTÉ
MICROFILMÉE TELLE QUE
NOUS L'AVONS REÇUE

Ottawa, Canada
K1A 0N4

NL 339 (Rev. 8/90)

THE UNIVERSITY OF ALBERTA

THE EFFECTS OF FEEDBACK ON TEST ACHIEVEMENT IN CAI

by

(C)

D. P. SHERIDAN

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY IN

EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

SPRING 1980

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled THE EFFECTS OF FEEDBACK ON TEST ACHIEVEMENT IN CAI submitted by D. P. SHERIDAN in partial fulfilment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

E. W. Ramanaula

Supervisor

Thomas O. Maguire

Raymond D. Schultz

Lauren Sandals

External Examiner

Date March 31, 1980

Abstract

This study examined the effects on long term memory of varying the following two common CAI constructs: (1) the timing of feedback message delivery, and (2) the content of the feedback message. The sample consisted of 60 graduate students enrolled in the 1979 Special Session, Faculty of Education, University of Alberta. They were registered in an 80 hour CAI course in introductory statistics. The data collection instrument was a 23 item multiple-choice test on the t-test. The test items were randomly ordered both as an end-of-chapter test and as a retention test seven days later. To determine if recognition or recall memory was in operation, several items were modified for the retention test.

The course and tests were presented using an IBM 1500 system. The research design required feedback to be provided either (1) immediately following a response to a test item, or (2) 24 hours later. The feedback message was either (1) a re-display of the multiple-choice item with the correct answer underlined, or (2) a series of sentences and/or a formula which provided a cue to the correct answer. Students were randomly assigned to one of eight cells in the research design. The first four cells were (1) immediate, correct answer feedback, (2) immediate, cue feedback, (3) 24 hour delay, correct answer feedback, and (4) 24 hour delay, cue feedback. The last four cells provided the same feedback timing and messages, but required the students to rank their

'certainty' of the correctness of their response to each test item, on a continuum from 1 (not certain) to 7 (absolutely certain). The IBM 1500 system collected all the students' responses, the time taken by students to respond to each item (response latency), and the elapsed time the feedback message was displayed on the terminal (feedback latency).

The results, in the form of mean test score data, indicated no difference between the immmediate and delay feedback groups on the retention test. No differences were found among groups on the recall or recognition test item scores. A significant difference in mean test scores existed between the correct answer feedback group and the feedback group on the end-of-chapter test, but not on the retention test. All groups scored better on the retention test in comparison with the end-of-chapter test scores. Requiring students to provide confidence values did not affect their test scores significantly. Also, the immediate feedback group did increase their mean confidence value between the test and retest. Analysis of the feedback latency data indicated no differences existed among the treatment groups, although the time taken to read feedback messages on the retention test declined significantly.

The study supports the current practice of providing a brief, corrective feedback message immediately following a response to a test item. However, delaying the time of feedback did not appear to have a deleterious effect upon

test scores. The nature of cues and their proper construction require more research. It was recommended that the impact of different feedback timing modes and variations in feedback message design be explored within the context of CAI by sampling from different age groups, with different levels of motivation, and using content from other subject matter areas.

Acknowledgements

I wish to thank the members of my thesis committee: Drs. E.W. Romaniuk, T.O. Maguire, A.E. Wall, R.A. Schultz, and M.W. Petruk, and Dr. L.H. Sandals who served as the external examiner, for their contribution to the completion of this study. My special thanks are expressed to Dr. E.W. Romaniuk, my thesis supervisor, for his guidance and supervision; and to Dr. T.O. Maguire for his encouragement and advice in data analysis.

I am also particularly grateful for the assistance and support given by all the personnel at the Division of Educational Research Services, in particular: Dr. S.M. Hunka, Mr. N.P. McGinnis, and Mr. J.E. Hunka . The working environment in the Division was one of unstinting helpfulness and infectious creativity. The numerous efforts of the computer operators and the willingness of the graduate students registered in STAT1 to participate in the study were much appreciated.

Finally, I acknowledge with special gratitude the faith and encouragement provided by my wife, Debbie, and my parents and family.

Table of Contents

Chapter		Page
I.	Background of the Study	1
	A. Introduction to the Problem	1
	B. The Problem and its Implications for Education	3
II.	Review of Literature: Memory Theory	7
	A. Introduction	7
	B. Accretion, Tuning and Restructuring	8
	C. Memory and Schemata	10
	D. Comprehension	12
	E. The Nature of Schemata	13
	F. Learning Through Accretion, Tuning and Restructuring	15
	Learning through Accretion	15
	Learning through Tuning	16
	Learning through Restructuring	16
	Memory and Learning	17
	G. Memory: Long Term Retrieval	18
III.	Review of Literature: Feedback	23
	A. Introduction	23
	B. Research on Feedback Timing	24
	Sturges	24
	Sassenrath	32
	More	35
	Kulhavy	36
	Surber and Anderson	39

Phye.....40
Newman, Williams and Hiller.....43
Summary of Feedback Timing Research.....45
C. Research on Feedback Messages.....48
Travers, Van Wagenen, Haygood and McCormick..60
Sassenrath.....61
Phye.....64
Kulhavy.....65
Summary of Feedback Message Research.....69
IV. Research Methodology.....74
A. Research Questions.....74
B. Design of the Study.....75
The Computer Assisted Course: STAT1.....75
The Instrument.....78
The Sample.....83
Instructions to Subjects.....85
C. Analysis of Data.....85
Feedback Timing Analysis.....86
Feedback Message Analysis.....97
Additional Analyses.....100
V. Conclusions and Recommendations.....116
A. Conclusions.....116
B. Recommendations.....120
References.....123
Appendix A. An Example of a Test Item.....134
Appendix B. Test Items and Cues.....139
Appendix C. Research Design.....152

List of Tables

Table	Description	Page
1	Recall & Recognition Test Scores	27
2	Seven Day Retention Test Scores	28
3	Summary of Sturges' (1972) Phase II Findings	29
4	The Phye(1970) Research Design	41
5	Impact of Feedback Messages	52
6	Sturges(1972) Phase I Findings	54
7	Summary of Sturges(1972) Findings	58
8	The Four Factor Experimental Design	83
9	Test and Retest Scores	88
10	Equal 'n' Test and Retest Score Means	89
11	Summary of 4-Way Analysis on Test Score Means	90
12	Merged Test and Retest Scores Means	90
13	Summary of 3-Way Analysis on Test Score Means	91
14	Mean Test Scores for Questions 19, 20, & 21	92
15	Summary Analysis for Questions 19, 20, & 21	93
16	Mean Confidence Values for All Responses	101
17	Analysis of Mean Confidence for All Groups	102
18	Mean Confidence Values for Correct Answers	103
19	Analysis of Correct Answer Confidence Values	103
20	Mean Confidence Values for Wrong Answers	104
21	Analysis of Wrong Answer Confidence Values	104
22	Feedback Latency by Treatment Group	107
23	Analysis of Feedback Latency	108
24	High Confidence and Feedback Latencies	109
25	Low Confidence and Feedback Latencies	111
26	Response Latency by Treatment Group	113
27	Analysis of Response Latency	113

List of Figures

Figure	Description	Page
1	A heuristic model of human information processing	15
2	A characterization of the retrieval process	19
3	Model of Sturges' (1972) research design	26
4	Forms of feedback messages (Sturges, 1972, Phase I)	51
5	Forms of feedback messages (Sturges, 1972, Phase II)	54
6	Relationship between feedback, confidence, and behavior	67
7	Mean test scores of treatment groups	87

I. Background of the Study

A. Introduction to the Problem

For centuries man has been concerned with memory or its antithesis -forgetting.

When Somonides offered to teach Themistocles the art of memory (450 BC) he is reported to have muttered wistfully, "I remember even those things which I would not, and can not forget what I would." Cicero observed, as have many others, "That memory is the treasury and guardian of all things."

As civilization advanced, philosophers mused over existence and nonexistence --memory and forgetting. Aristotle apparently believed the invention of writing would cause memory to lose its facility and gradually disappear. Similarly, a Chinese proverb of the period states: "A clever memory is not equal to a clumsy brush". Contemporary commentators might suggest the photocopier could do the same.

Man has been preoccupied with attempting to remember many kinds of data over the centuries: from the size of an elephant herd, to the text of a message memorized and carried by a courier, to a current concern for telephone numbers, household addresses, clients' names, financial accounts, credit card expenditures and so on.

Within education, attention has been drawn to recall scientific formulae, historical events, literary style, and mathematical models or algorithms. Most teachers have

experienced the amazement of a student's failure to remember material when several weeks, days or hours before that same student demonstrated some competancy with the material. Knowing, and apparently later not knowing, causes educators and students alike to examine their teaching/learning strategies and curricular approaches. Massed versus distributed practice, reviews, short quizzes, spiral curricula, and advanced organizers are only a few of the assaults on the apparent insidious degradation of memory. Although hundreds of studies have been done, solutions to the problem of retention are too few.

Cermak commented:

> ... the application of memory research to education is upon us this century. Education's demands must be met. Psychology has hidden its head in the sands of irrelevancy for too long. It must be held responsible for the application of some of its findings to the education of children... Psychology has been loth to apply any findings to education because the educators do not understand the theories behind them, and educators have not let psychologists experiment in the classroom because the psychologists do not understand the basic processes of educating humans.
>
> In memory research, humans, not rats, are being investigated and it is time for the parties to realize this. It is admittedly a gamble when a new method is used in teaching, but it is going to be necessary. In the future, as more is discovered about memory it must find its way into the classroom; it must be useful; it must be applied. (Cermak, 1972, p.268)

In the Second Handbook of Research on Teaching Glaser (1973) concluded that research on instruction in the schools has proceeded at a snail's pace. He stated this is partly the result of the difficulty in adequately controlling the variables or processes involved. However:

> The computer now makes it possible to have instructional procedures selected systematically and the resultant learning observed in the school context. (p.851)

In a later passage he commented:

> The positive potential of educational technology will only be realized if the technologist who would bring the results of their science to bear on the educational problems are actually concerned with the broad goals of education and make a concerted effort to fully assess the effects of that new technology. (p.856)

Kulhavy (1977) in a review of studies examining feedback and written instruction observed, with specific reference to CAI, that "because computerized instruction allows such a wide range of strategies for each response, the question of how one most effectively matches feedback parameters with response characteristics is indeed an important one." (Kulhavy,1977)

This study examines the application of memory theories and learning research to instructional design using the new educational technology, CAI. The next section presents a description of the research problem and the implication solutions to this problem have for education.

B. The Problem and its Implications for Education

Computer-assisted instruction is a new technology which blends instructional courseware with digital computer devices to provide an instructionally consistent interactive learning environment. In a tutorial mode, for example, it is usual for subjects to be immediately provided with feedback messages. Those messages may range from a brief sentence

stating that the student's response was right or wrong to a more detailed remedial paragraph filling the screen of the computer terminal.

Currently, in CAI, a continuing widespread notion is one which suggests feedback should be immediately provided (since it is possible under CAI and not normally possible under conventional instruction) and another is that feedback should be brief, and corrective or reinforcing. To date, very little research evidence has been accumulated to confirm or reject the validity of these assumptions. This study assesses these questions.

Instructional designers recognize the interrelationship of learning with memory. The design of learning activities involves an assessment or assumption of previous knowledge or skills, an activity component focused upon developing or adding to these skills or knowledge, and an assessment stage to measure the success of the activity. This would then be followed by remediation and retesting, or movement to a new learning objective. There is a two-fold need to have subjects quickly and effectively achieve the instructional objective as well as to retain the knowledge for subsequent use and as a building block for future growth. The problem for CAI authors is one of selecting a learning strategy which has a high probability of providing both good short term success and good long term retention. It is believed learning designs must provide environments and strategies most likely to produce long term retention if overall

learning activities are to be worthwhile.

This study examined two commonly used instructional design constructs specifically for their effects on long term retention. The questions asked were:

1. Does immediate feedback result in better long term retention than feedback delayed 24 hours?
2. Does a feedback message which consists of underlining the correct answer in a multiple choice question result in better long term retention than a feedback message which consists of a cue to the correct answer?

In addition to the delivery of instruction, the IBM 1500 system was used in this study as an important data collection device. These data were used to satisfy additional instructional questions. In brief, the IBM 1500 CAI system includes a student performance accounting program which supplies a record of every student's response, the location in the course associated with the response, and a measurement of the time taken for the response to be entered (in tenths of a second). As a result, an instructor can ascertain: if all students have covered specified material in the course, that all have been tested in the same manner, and that precise records exist to describe their activity.

In this study two variables have been under examination: (1) feedback timing (immediate and 24 hour delay), and (2) type of feedback message (an underlined correct answer, a cue to the correct answer). With respect to these two variables, and utilizing the IBM 1500 student

performance accounting programme, supplementary questions were examined to determine if these two variables affect:

a. the mean confidence that students assign to their responses,

b. the mean latency time that subjects require to produce responses, and

c. the mean latency time that subjects take to read a feedback message.

The review of related literature follows in Chapters II and III. Chapter II provides an overview of an information processing theory of memory, a theory which is a useful model to explain the research findings of Chapter III. Chapter III examines a number of studies which indicate that the use of immediate feedback and brief messages in conventional instruction may not always result in long term retention by students.

II. Review of Literature: Memory Theory

A. Introduction

The initial impetus for this study arose from the work of P.T. Sturges, a long time researcher in the area of feedback and retention. Her work, reviewed in depth later, may be characterized as a series of investigations of long term memory in which the experimental design is systematically modified and fine tuned. So far as can be determined, Sturges has not placed her findings in any particular theoretical camp, nor has she debated at length the broad theoretical implications of her findings. As a result, Norman's theory of learning and memory is reviewed, to provide a theoretical structure to explain those research findings which indicate (1) delay in feedback may improve retention, and (2) a feedback message which is a cue to the right answer may also improve retention.

Attention is first directed to a theory of learning and memory. Norman points out:

> The study of learning differs from the study of memory in its emphasis, not necessarily in content. Learning and memory are intimately intertwined, and it is not possible to understand one without understanding the other. (Norman, 1977, p. 1)

Norman's theory belongs to the school of semantic memory, one which addresses itself in a somewhat phenomenological way to the content of an individual's memory, i.e., the characteristic acquisition and use of information. Norman differs from many semanticists by attacking what he perceives to be a weakness in the semantic

school; that is, the disuse of the term "learning" in favor of a process he describes simply as the "acquisition of information" in memory (Norman, 1977, p.1). The thesis that evidence of learning is demonstrated by the ability to retrieve appropriate data on cue is found trite. Norman argues the simplicity of this theory is challenged by the emergent quality of the retrieval; one which includes not only the encoding and processing during input but appears to have involved a merging of information collected over time or the development of new forms/structures for the current data.

In complex learning there is what may be characterized as an insight, a "click", or an "ah ha". It is this internal operation which is placed under scrutiny by Norman and has lead to the theory he terms the "Active Structural Network of Long Term Memory". His goal was to establish a general integrated theory capable of describing systems that acquire, interpret and use information.

The following section describes Norman's theory of memory and relates this theory specifically to learning.

B. Accretion, Tuning and Restructuring

Three quantitatively different modes of learning are proposed - accretion, tuning and restructuring.

Learning through accretion is perceived to be the daily accumulation of information, a process of acquiring facts, lists, names, numbers, and so on. This knowledge accumulates

and increments data bases in an unsophisticated manner. Norman suggests no structural changes occur in the information processing system itself and that accretion is the type of learning most studied by psychologists.

Learning through tuning is not only the accretion of information but the process of changing the criteria used for processing the information. Schema - dynamic processing units - normally used for sorting and storing data in the accretion mode are under the tuning mode modified to bring themselves into congruence with the functional demands placed upon them.

> Thus, for example, when we first learn to type we develop a set of response routines to carry out the task. As we become an increasingly better typist these response routines become tuned to the task and we become better able to perform it more easily and effectively. (Norman, 1977, p. 4)

A child's increasing specificity - from the classification of all small four legged animals as "doggies" to the genus/species to which the animal best belongs - may also be an example of the tuning of schemata. Similarly an adult's conclusion that all light aircraft are "Cessnas" is modified as knowledge of light aircraft design increases.

Learning through restructuring is a more significant and different process that occurs when new schema are required to interpret new information or reorganize what has been acquired. Restructuring leads to efficiencies in retrieval, interpretation and acquisition of new knowledge. It is suggested only an inner sense of the "unweildiness or unformedness of the accumulated knowledge gives rise to the

need for restructuring" (Norman, 1977, p. 4). Accretion and tuning occur continually. Restructuring may take days, weeks or years depending upon the nature and flow of the information and the critical mass needed to cause a resorting of the data, reformatting of schema (tuning) or creation of new schema to process data parsimoniously. Examples of students and athletes who show a growth in skills over many years are examples of the restructuring phenomena. Fitts commented:

> The fact that performance can level off at all appears to be due as much to the effects of physiological aging and/or loss of motivation as to the reaching of a true asymptote or limit in capacity for further improvement.
>
> (Fitt, 1964, p. 268)

In summary, accretion is the process of data classification and storage, tuning involves the modification of schema to insure better accretion, and restructuring occurs when current schema no longer appear adequate to the data base and thus new memory structures are needed. It is memory schema in general that is of concern in this study.

C. Memory and Schemata

Memory may be considered as specific or general in nature. Specific retrieval deals with such things as what occurred at 10:00 Monday morning as compared with, "What do you think of Joe Brown?", which is a composite of many specificities. Other examples may be the characteristics assigned to the schema "dog" or the schema "farm". Norman

states:

> To us, a schema is the primary meaning and processing unit of the human information processing system. We view schemata as active, interrelated knowledge structures, actively engaged in comprehension of arriving information, guiding the execution of processing operators. In general, a schema consists of a network of interrelations among its constituent parts, which themselves are other schemata. (Norman, 1977, p. 7)

Within schemata are variables which are "references to general classes of concepts that can actually be substituted for the variables in determining the implications of the schema for any particular situation" (Norman, 1977, p. 7). As information accrues it is encoded against or substituted for the variables of a general schema. This memory will thereby become a specific, particularized, or an instantiation of the general schema. An example of a general schema is "automobile" or "car". The "car" will be represented by a highly detailed schema by a mechanic and to a lesser and much modified extent by an operator, used vehicle salesman, or potential purchaser. It is often astonishing to learn of someone who is apparently indifferent to fluid levels (e.g., crankcase, battery, radiator) yet highly sensitive to such variables as the car's color, upholstery and carpet. Clearly schema are individual in character. As a result two persons may view the same automobile, encode and process information about it and later retrieve facts in a seemingly integrated fashion and yet retrieve data using a different organization and recall both similar and unique attributes. Indeed some

salient features may not be retrieved at all. It is also not uncommon to retrieve more data than were initially available since assumed variables may carry over from the general schema. For example, the assessment of a used car by a potential buyer --the assumptions made prior to purchase-- will become only too obvious with time!

Variables may also be defaulted or constrained. The variables of a general schema can have values assigned to them by default or restricted by a range of possibilities. In the purchase of a house, the uninformed buyer, upon viewing the estate, may believe fixtures come with the house or that a specific fixture will remain. Experience results in tuning or restructuring the general schema to reduce the default or constraining process.

> Variables (and their constraints) serve two important functions:
>
> 1. They specify what the range of objects is that can fill the positions of the various variables; and
> 2. When specific information about the variable is not available, it is possible to make good guesses about the possible value.
>
> (Norman, 1977, p. 10)

D. Comprehension

Comprehension may be confirmed when retrievals indicate an appropriate configuration of schema have been used to account for an event or situation. This implies that the composition of each schema has identified the salient concepts and events within each occurrence.

The process of comprehension, therefore, involves verifying or rejecting various schema until some level of harmony/coherence is achieved. As a result of processing (restructuring) new schema may emerge with new bridging between variables. Efficiencies will appear in the interpretation of the data base, searching and retrieving, variables and processing of new information.

Like a good theory, schema account for existing facts through a parsimonious description of a universe and has potential to accomodate new discoveries. A schema is created to explain or describe a situation and remains unchanged even with substantial growth in the data base, so long as its utility for encoding and retrieval remain valid. Once it becomes clear that a schema will no longer support stored data, a process of either tuning or restructuring the schema occurs. If an insufficient data base exists for the creation of a schema then the information may remain for a time as disconnected subsituations, each interpreted in terms of a separate micro-like schema, for example, a hitherto unrelated fact (nonsense syllable).

E. The Nature of Schemata

Norman considers " schemata as active processing units, each schema having the processing capability to examine whatever new data are being processed by the perceptual systems and to reorganize data that might be relevant to themselves" (Norman, 1977, p. 11). Schemata, activated when

appropriate data appear, guide data organization according to their structure. Schemata control and direct the comprehension process itself. In addition, it is suggested that output from one schema may reenter the data stream to become input for another schema. Reddy uses the image of a blackboard to explain the phenomena (Reddy and Newell, 1974). Data may be thought of as appearing on a blackboard to be examined by relevant schemata cued by the nature of the material. Data relevant to a particular schemata are processed using internal conceptualizations and rewritten on the board. (Figure 1 illustrates the process. The blackboard may be considered as existing in the synthesis/interpretive space.) This modified information may cue other schemata which in turn process and redisplay their output. A halt occurs when schema are no longer cued by data on the board (stream). Naturally the cyclic result will be a reflection of the efficiencies of the schema and their convergence with reality as tested on subsequent instances. The schema-data cycle has been discussed under such headings as: "active demons" (Selfridge and Neisser, 1960), "actors" (Hewitt, Bishop and Steiger, 1973), and "production systems" (Newell, 1973).

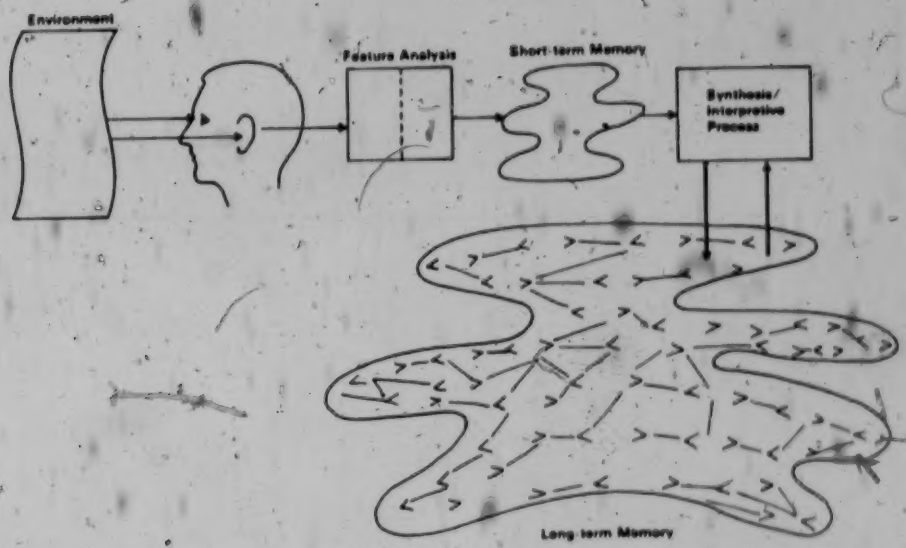


Figure 1. A heuristic model of human information processing

F. Learning Through Accretion, Tuning and Restructuring

Learning through Accretion

As discussed earlier, the main mode of learning is simply the daily absorption of information. Norman describes instantiations as newly created data structures patterned on old schema but carrying current information in place of the variables. This representation of an event is placed in long term memory and retrieved using the general schema to reconstruct the earlier, original experience. This remembrance is a process similar to storage and is activated using similar schema. Learning by accretion therefore

processes all information in a similar fashion, i.e., with the same schema. If data can not be configured using current schema then tuning/restructuring must occur before learning may proceed, otherwise the data remains as substructured and with independent micro schema unconnected to other data piles.

Learning through Tuning

For Norman and his associates, the modification of existing schemata to better process and store data is a matter of "fine tuning" the structure. Basically, revising constant and variable terms has the effect of:

1. improving the accuracy of analysis of information,
2. generalizing the range of applicability (replacing a constant with a variable or modification of a variable),
3. specializing the applicability (constraining variables or replacing a variable with a constant), and
4. determining the default values (discovering the attributes which normally apply and adding these to the schema such that intelligent guesses may forward inference making and guide further processing).

Learning through Restructuring

So long as existing memory structures adequately account for new knowledge, tuning and restructuring are not required. In a typical learning situation, information would be accreted until the body of knowledge becomes unmanageable through poor or inaccurate retrieval. At this point either new or tuned schema are created that enhance processing and

improve retrieval (evidence of memory). Norman suggests two types of schema creation occur: pattern generation and schema induction.

Under pattern generation a schema is copied, then modified as required. Learning through analogies is an example of this process, e.g., learning that a rhombus is to a square what a parallelogram is to a rectangle. The constants from one schema are modified in the new one. Learning to differentiate breeds of dog also indicate restructuring.

Schema induction, the other form of learning, results from either a spatial or temporal co-occurance which cue several schema. This simultaneous activity or temporal contiguity

> ...is the fundamental principle of most theories of learning, but it seems to have amazingly little application in the learning of complex material. As far as we can determine most complex concepts are learned because the instructor either explicitly introduces an appropriate analogy, metaphor or model ... We believe that most learning through the creation of new schemata takes place through patterned generation, not through schema induction. (Norman, p.16)

Memory and Learning

Incoming data are most efficiently processed when they are consistent with existing schemata. The more that the arriving information deviates from a person's current interpretive structures, the greater the need for change either through tuning or restructuring. However, this presumes a recognition of discrepencies. If through misinterpretation or misunderstanding the material appears

consistent with previously processed data, the need for change will not occur.

> Reorganization of the memory system is not something that should be accomplished lightly. The new structure that should be formed is not easy to determine: the entire literature on "insightful" learning and problem solving, on creativity, on discovery learning, etc., can probably be considered to be studies of how new schema get created. We do not believe that the human memory system simply reorganizes itself whenever new patterns are discovered: the discovery of patterns, the matching analogous schemata to the current situation most probably require considerable analysis. (Norman, p. 22)

G. Memory: Long Term Retrieval

Williams (1977) characterized the act of retrieval from long term memory as a reconstructive process. The operation, as he sees it, involves a recursive cycle of three phases which switch alternately from (1) finding a context, (2) searching, and (3) verifying.

The operation commences with a sketchy description (which Williams terms a context), and a search begins. If anything is found it is immediately tested against what is known or proposed, i.e. the process of verification. If this fact is accepted the objective may have been satisfied; if not, the new fact is added to the description and the search then proceeds one level lower (in the recursive sense) using a more definitive context. An outcome of Williams' (1977) theory of retrieval is the assessment of confidence, which is expressed as the degree of certainty the individual assigns to the retrieval, i.e., confidence levels.

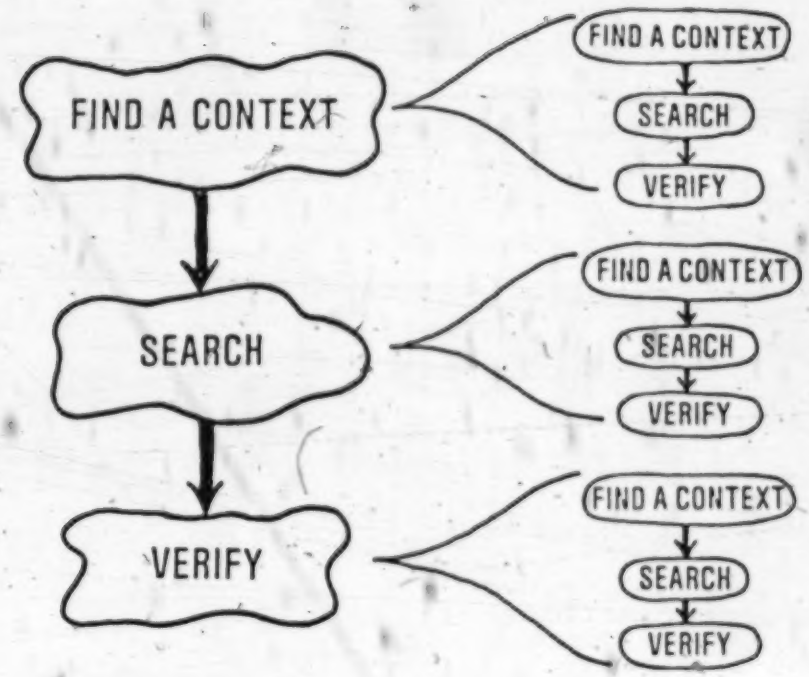


Figure 2. A characterization of the retrieval process.

In order to better understand the search process, Williams (1977) proposes two metaphors. One is the suggestion that the individual continues to aggregate information so that he "homes in on the target". The flow of information continues to build a more and more precise description. The search process starts with a generic context that iteratively becomes more specific.

A second metaphor is that of the "jigsaw puzzle" suggesting an algorithm which narrows the search field to a region that looks most promising (working on the border of a

puzzle) and builds up a better formed context before proceeding into areas that require the assessment and inference on a larger scale with less likelihood of solid verification.

The size of the search description may cause problems in verification. Too little information makes the verification process weak since there may be:

1. a large number of possible events;
2. recall of a re-encoded event is more likely than the original (fewer more apparently typical properties); and
3. if a searching property is not available then knowledge indexed under that property must be inferred.

The problem of too much information also makes verification difficult. The selection of contexts to determine the correct one results, if a bad choice is made, in a nil retrieval and the report "I forget". Verification constraints ideally should result in dropping the search property and returning to the context so far created to investigate and test other possible leads.

It is often reported, however, that a property, verified and found false, continues to obstruct the process. These "distractors" may appear to totally frustrate an individual who, while trying to remember something, may report the continued reappearance of a particular fact which is impeding his progress. Often simply examining the distractor so as to become totally conscious of its existence and then consciously discarding it succeeds in

"pigeon holing" it and thus removing it from active consideration. Clearly the distraction has some connection with the context being constructed and is sufficiently powerful to cause aggravation. A distractor is therefore identified through the process of verification, as an interference with the retrieval process, which in a limited domain matches the item sought.

Confidence, when used in connection with retrieval, is a statement about the results of a verification. Three techniques of truth testing are:

1. Coincident recovery - discovery of a similar fact from another source.
2. Indirect confirmation - a retroactive verification arising from information uncovered as a result of using a previous unconfirmed description.
3. Consistency checking - information fits what is already known and it is therefore considered correct.

The extent to which some or all of these truth testing conditions occur and are satisfied is directly reflected in the degree of certainty assigned to the retrieval.

Summary

From this characterization of memory by Norman and Williams, one is provided with an interpretation of memory phenomena. The three stage recursive nature of context seeking, searching, and verification function upon a base of partial information and descriptions to reconstruct

previously encoded, stored data. A description is used to seek fragments of information which, if verified, are added to the description to retrieve still more information until a match on the material sought is made. The condition reported as "forgetting" is considered under a number of possible headings: too little information, too much information, false recoveries, and re-encoding. Generally, failure to retrieve is caused by building a search descriptor (composed of many search terms) for which no schema exist. The result is a null retrieval. In some instances, the search terms will be valid and the individual will sense that some form of a memory is in existence, yet report he has forgotten or perhaps it is on the 'tip of his tongue'. Williams (1977) also indicates that subjects are capable of verifying retrievals and ascribing a degree of confidence to the data retrieved.

In the following section, Chapter III, the literature reviewed highlights findings in support of delaying feedback delivery and designing feedback messages to provide more than just the correct answers. Chapter III begins with a review of the effect of delayed feedback upon long term retention.

III. Review of Literature: Feedback

A. Introduction

Early investigations of feedback begin with Judd (1905) and his study of practice without providing knowledge of results. In the decades following, emphasis shifted away from the study of "academic" learning and feedback to the study of psychomotor activity and feedback. The reason for this shift was in part due to the United States military funding of research which was directed toward the investigation of methods for the improvement of training (psychomotor) programmes. A summary of these studies was provided by Ammons (1956). Interest in feedback was rekindled when it was found that feedback could be manipulated to produce differential results in long term retention. It was Brackbill, et al., (1962a, 1962b, 1964a, 1964b) who discovered and termed this phenomena the delay-retention effect (DRE). On cognitive tasks, delaying informative feedback by as little as 10 seconds produced better retention many days later. Other dedicated reserchers of DRE are Sturges (1969) and Sassenrath (1968). The work of Sturges is reviewed first because of its historical precedence, comprehensiveness, and contribution to understanding DRE through continuing investigation. Other papers are discussed which reexamine or confirm DRE.

Generally feedback researchers have not placed their findings in the context of a psychological theory, perhaps because earlier classical theorists (Skinnerian) could not

explain DRE. The human information processing theory, a more recent model which describes the multidimensionality of memory, appears useful as an explanation of the effects researchers have found as a result of varying feedback timing and message design. It is believed that the preceding section on memory theory provides a theoretical framework helpful in understanding the work of the following researchers.

This chapter is organized in two sections, the first reviews the literature concerned with feedback timing and its effects; the second surveys the research related to feedback message design. It will be noted that some researchers have examined both feedback timing and feedback message design within the same paper. In these cases the paper will be discussed separately. It is believed this organization of the data will better aid in evaluating the two bodies of research. The section which now follows considers the research on feedback timing and begins with the papers by Sturges.

B. Research on Feedback Timing

Sturges

In her initial study, Sturges (1964) examined the effects of immediate feedback and 24 hour delayed feedback upon long term retention. The content for this study, which consisted of a combination of uncommon English words as well as some nonsense material, was presented in the form of a

multiple choice (M/C) test with items displayed using a 35mm slide projected on a screen. University level subjects were tested on an individual basis. The time available for responding to the test item and reading the feedback message was fixed to a number of seconds (details not available).

Seven days later, on a retention test of the material, subjects who received the delay feedback were found to have significantly higher scores on meaningful material than those who received immediate feedback. Sturges found that the variation of time allowed for feedback had no apparent impact upon retention of nonsense material.

Sturges (1969) reconfirmed the findings of her 1964 study using test material from the social sciences area. The time required for responding to the test item and reading the feedback message was again controlled. In this case, responding to the item was restricted to 20 seconds and feedback was presented for 10 seconds. Sturges concluded that 24 hour delay in feedback was superior to immediate feedback, but added the proviso that the effect could be neutralized through a manipulation of the form of the feedback message. This study is reported in greater depth in part 2 of this chapter.

Sturges (1972a, 1972b) examined the effect of providing feedback immediately after a response, at the end of test (EOT), and 24 hours following the test. Retention testing took place immediately after feedback, or not at all (control group) and seven days later for both groups as a

final retention measure. Retention test items were designed to test both recall and recognition. The subject matter was based upon uncommon English words and the response and feedback latency times were standardized at 15 seconds.

The procedure followed by Sturges (1972) consisted of a testing algorithm with (1) three delay modes, (2) three immediate tests (nil, recall, recognition) for practice, and (3) after seven days a recall/recognition test. Figure 3 provides a schematic of this research design.

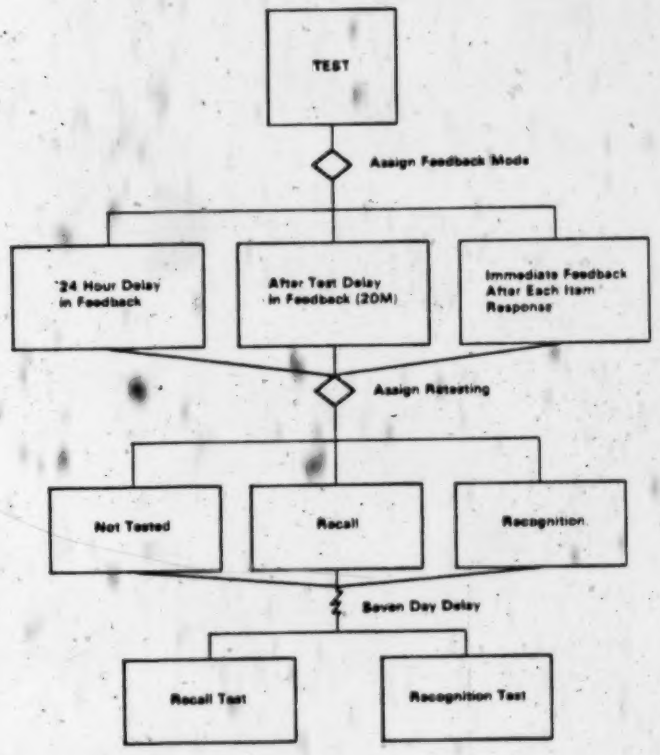


Figure 3. Model of Sturges (1972) research design

The findings on immediate retesting following feedback, summarized in Table 1, were as follows:

1. A significant difference in test scores ($p<.001$) existed between the delay in feedback group and the group receiving immediate feedback.
2. There was a significant interaction between the delay and form of test ($p<.01$). Recall was enhanced to a greater degree by delay than was recognition.

Table 1
A Comparison of Recall and Recognition Test Scores Immediately Following Feedback

	Zero Delay	EOT-24H Delay
Recall	6.62	11.19
Recognition	26.96	28.26*

*Significant at the .001 level

The findings on the seven day retention test were as follows:

1. Feedback delay groups performed significantly better ($p<.01$) on recall items than on recognition items.
2. Overall recognition test scores were significantly better ($p<.001$) than recall test scores.
3. Those not receiving immediate testing scored significantly lower ($p<.001$) than those tested immediately.
4. Immediate recognition testing lead to significantly better retention ($p<.001$) than did immediate recall testing.
5. A significant interaction was observed between the

immediate test form and the seven day test form ($p<.001$). The immediate recognition test group scored higher on the retention test recognition items than on the retention test recall items. Immediate recall testing contributed more to retention test recall scores than did immediate recognition testing. Thus, recall testing improved recall while recognition testing improved recognition.

Table 2 provides a summary of the significant results.

Table 2
A Summary of Seven Day Retention Test Results

Group Scores		Group Scores
EOT-24H Feedback Delay	>	Zero Feedback Delay***
24H Feedback Delay	>	EOT Feedback Delay**
EOT-24H Feedback Delay on Recognition Test	>	EOT-24H Feedback Delay on Recall Test**
Overall Recognition	>	Overall Recall***
Immediate Test	>	No Immediate Test***
Immediate Recognition Test	>	Immediate Recall Test***
Immediate Recognition Test + Seven Day Recognition Test	>	Immediate Recognition Test + Seven Day Recall Test***

*Significant at .05 level
**Significant at .01 level
***Significant at .001 level

On the strength of this study (Sturges, 1972a, Phase I) a prescription might read:

a. If maximum seven day recall is desired, then delay feedback for 24 hours and then immediately administer a retest using a recall question format.

b. If maximum seven day recognition is desired, then delay feedback for 24 hours and then immediately

administer a retest using a recognition question format.

In the second part of the study, Sturges (1972a) varied the feedback message design and again contrasted the effects of immediate feedback, end of test feedback, and 24 hour delay of feedback. Delay was again found superior to no delay, although no significant difference was found between the end of test feedback group and the 24 hour delay feedback group. Immediate retesting following feedback significantly increased seven day retention scores. Again, immediate recall testing enhanced seven day recall scores and immediate recognition testing enhanced seven day recognition test scores.

These findings are summarized in Table 3.

Table 3
Summary of Sturges (1972) Phase II Findings

Immediate Test Results:		
Recognition Scores	>	Recall Scores***
Retention Test Results:		
Delay Group Scores	>	Immediate Feedback Group Scores**
Recognition Scores	>	Recall Scores***
Scores if Immediate Retesting Occurred	>	Scores without Immediate Retesting***

**Significant at .01 level
***Significant at .001 level

Sturges (1972a) concluded that superior retention with 24 hour delay of feedback was due to factors operating at feedback, not factors intervening between the test and retest.

> These findings support the interpretation that the delay retention effect depends upon: (a) stimuli present during feedback, (b) how the subjects respond to these, and (c) relevance of these stimuli and responses to the retention test. (Sturges 1972a:41)

A delay in feedback removes the subjects from the immediate concern of "Was I right or wrong?", to a more encompassing appraisal of available data and interrelationships. Sturges indicated that the effects of immediate feedback could be enhanced by an immediate retesting at the end of the test. It was pointed out that the objective of feedback is to cause subjects to understand the test material better and thereby improve on the next test.

Sturges (1972b) primarily manipulated message design to determine which feedback construction resulted in the best long term retention. Zero, end of test, and 24 hour delay feedback modes were also a part of this research design. The findings again confirmed the superior impact that delay of feedback has upon long term retention scores when using certain types of feedback messages.

Sturges (1976), examined feedback under computer-managed testing (CMT) and detected the same effect

for delay as did her previous research. This study compared the differences among feedback modes which were given immediately after each item, at the end of test (EOT), and 24 hours after the test. The dependent variable was the set of scores on a retention test given one to three weeks later. Subject matter for the test items (30 M/C) was drawn from a University of Calfornia child psychology course and administered via cathode ray tube computer terminals connected to a PDP 11/45. The programming language was SOCRATES. Retention was measured using a 47 item criterion test composed of 30 previous M/C items plus 17 additional short answer items. Student anxiety state and confidence were also measured. The A-State anxiety scale was administered before and after the computer managed test (CMT) and the retention test. Confidence measures were solicited following each item of the CMT and retention tests.

The findings support the use of feedback over no feedback and delay (EOT, 24H) over no delay. In fact, the longer delay positively affected the confidence students had in their answers. The number of items that were wrong on the immediate test but right on retention test confirm the value of delay of feedback. On items judged most difficult, retention scores indicated that improvement was related to the feedback conditions. Evidence of the influence of delay was the increased confidence students indicated when completing the retention test. The 24H group

was more confident than all other groups and the EOT was more confident than their counterparts who received immediate feedback. Anxiety findings were inconclusive.

Sturges concluded:

> Long term retention of academic material following immediate feedback is not superior to that with delayed informative feedback. Some delay in presentation of informative feedback results in superior retention performance, and the longer 24H delay is superior on the measure of confidence ratings. (1976 p.ii)

In the NPRDC report Sturges (1978) recommended:

> 1. Further study should be conducted to extend the findings of the present study by comparing the relative effects of immediate and delayed feedback under other experimental conditions (e.g. using different forms of feedback presentation and/or criterion test items and conducting repeated computer managed tests with informative feedback throughout a course).
>
> 2. It is assumed that these results are due to an increase in student concentration on feedback that influences the level or breadth of processing of the remembered information and the feedback. Therefore, procedures that foster the breadth of processing should be developed and evaluated.

Throughout the period of Sturges' activity in this field, other researchers have also been actively examining the effect of delay upon long term retention.

Sassenrath

Sassenrath and Yonge (1968) also used meaningful material to examine the effect of feedback delay upon retention. In this instance, the contrast was simply between (1) immediate feedback in serial fashion after the subject responded to all test items (EOT) or(2) delaying 24 hours

before presenting the feedback. On a seven day retest, the group which experienced a 24 hour delay before feedback achieved scores significantly higher than those receiving immediate feedback.

Sassenrath (1968) examined the delay-retention effect using a 60 item, four alternative, M/C test based on introductory psychology. Students were alotted 15 seconds to read an item and to record an answer on an IBM answer sheet. After the test, one-half of the immediate informative feedback group received a copy of the quiz items (stem and answer options) with the correct answer underlined. The other half of the immediate feedback group received only the answer options with the correct answer underlined. Within each group, one-half were informed that they should try to remember the answer as a retention test would follow. Ten seconds were allowed for Ss to read each item. The same protocol was followed with the delayed feedback group, but 24 hours after the initial quiz. Immediately following feedback, and five days following feedback, the groups were retested using the same items in random order. These retention tests were written at the students' pace.

Findings indicated no significant difference between feedback groups on the initial test or the first retention test (immediately after feedback), yet after a five day interval a significant difference ($p < .001$) in favor of the delayed feedback group appeared. Those receiving the message to learn and retain the answers because of a future test

performed significantly better than the group which received only the alternatives with the correct answer underlined.

Sassenrath concluded:

> There is mounting evidence that delayed informative feedback does not retard learning and may enhance delayed retention If so, these results have considerable implications for learning theory, programmed instruction and classroom teaching. (p. 72)

In an elaboration of the previous study, Sassenrath (1969) decided (1) to test the effect on each item of varying delay of feedback between immediate (one second) and delayed (ten seconds), and (2) to provide four types of feedback (discussed in detail in part 2 of this chapter). Retention was measured both immediately after feedback and five days later. The Ss were 311 upper year college students and the M/C (four alternative) test was based on introductory psychology. The procedure was to randomly assign Ss to the treatment groups and administer the test. Item presentation and feedback was via slide projector in a group setting. Fifteen seconds were alotted for answering each item on a IBM answer sheet. Following the response, either 1 second or 10 seconds passed before feedback was presented. After answering the last test item and viewing the feedback sequence, the test was readministered. This time the students progressed through the items at their own pace. The retention test was administered five days later.

Once again, the delayed feedback group performed significantly better ($p < .05$) on the retention test (five days later), although there was no significant difference

detected between groups on either the initial test or the immediate retest following feedback. Sassenrath noted:

> Although the differences in retention are usually not large in absolute amount, the psychological importance of the difference contrary to the accepted principle that immediate reinforcement, as opposed to delayed reinforcement, produces superior learning and, therefore, presumably superior retention. (p. 176)

More

More's (1969) mammoth work on feedback delay and retention examined the effect of four different delay periods (none, 2.5 hours, 1 day, and 4 days) on retention. This was measured three days following feedback. The subjects were 663 grade eight students who read two articles of 1200 words each on the topics of glaciers (science) and Rhodesia (social studies). A 20 item M/C test followed. Immediate feedback was provided using erasable answer sheets. Delayed feedback was provided by returning to test booklets to the students. The test booklets contained strips glued to the right hand margin of each page. Each strip indicated the correct letter responses to the questions on the page. Retesting indicated that a delay of between 2.5 hours and one day produced optimal scores. An acquisition criterion group was tested immediately following each of the four delay modes. Highest scores were obtained for those experiencing the 2.5 hour and 1 day delays. Thus, the optimal delay not only provided information that resulted in better marks immediately following delivery but the benefit carried over to the retention test.

More argued that the primary objective of instruction and testing is the retention of what is learned. His study indicates the need to time the return of graded tests appropriately. To do otherwise "may not only be ineffective, but may actually inhibit retention learning." (p. 342)

Kulhavy

Over the past seven years Kulhavy and R.C. Anderson have coauthored several papers which investigated the delay-retention effect (DRE). They later identified and explored the phenomena termed interference-perseveration. From the beginning Kulhavy attempted to explain DRE and to demonstrate its existence.

> According to the interference-perseveration hypothesis, when a person makes an error on the first test, he strengthens an A-B connection which then interferes with acquiring an A-C connection from the feedback. Proactive interference is greatest when stimuli in successive tasks are identical and the responses are dissimilar. This, it is argued, is the condition that prevails when an incorrect response is made on the first test. According to this analysis, a person who makes a correct response choice on the first test places himself in the A-C A-C paradigm, a condition known to facilitate retention. (Kulhavy & Anderson, 1972, p. 507)

Kulhavy cites in support Anderson & Myrow (1971), Roderick & Anderson (1968), Rothkopf (1966), and Spitzer (1939). These studies consistently demonstrated that tests following instruction consolidated learning so that performance was improved on successive tests. This improvement through testing was independent of feedback.

Evidence that errors perseverate after an initial test

was presented by Kaess & Zeaman (1960). By manipulating the number of incorrect alternatives on a M/C test, they detected the continuation of these errors on succeeding tests. Only after several trials were subjects able to abandon their earlier performance.

Built into Kulhavy's interference-perseveration hypothesis is the notion that time is a critical determinant in the success of feedback. In support he noted that delay has been found to reduce proactive interference in non-academic learning (Abra, 1969; Underwood & Ekstrand, 1967; Underwood & Freund, 1968). Kulhavy also addressed the problem of attention at immediate feedback (a notion first put forward by Sturges, 1964). That is, because of frustration and fatigue, the learner does not process the presented data as carefully as he should. Kulhavy concluded that an analysis of time spent observing feedback should be evidence of the processing occurring, irrespective of the delay or non-delay in feedback.

Kulhavy & Anderson (1972), employed 194 high school students taking introductory psychology. A printed booklet containing a 35 item M/C test was administered. Feedback was given by returning the test booklet with correct answers underlined. Test and feedback items were randomized for each subject. A feedback delay of 24 hours versus immediate feedback was in effect. A second (retention) test was administered one week following the first test. The time, to the nearest minute, taken to read the feedback booklets was

recorded by the experimenter. Subjects were not informed of future tests. The total time allowed for test and feedback was one hour. The learners working in a self-paced manner completed the tasks in the allotted time.

Kulhavy found that the groups receiving delayed feedback were significantly different ($p < .01$) from those receiving immediate feedback when compared according to the probability of proportion of answers wrong on the second (retention) test as compared with those items wrong on the initial test. A significant difference ($p < .01$) also existed between the immediate and delay feedback groups when compared on the basis of test scores. If the theory of reinforcement applies, it should be expected that the ratio of Right2(retention):Right1(initial) should be higher for the immediate feedback group than for the delayed feedback group. However, the data did not support this theory.

An unusual, additional facet of the feedback puzzle was explored in this paper. Subjects were requested to identify their previous errors when they received feedback. Not surprisingly, those in the immediate feedback group identified significantly more errors ($p < .01$) than did the delayed group. Thus, more forgetting in the delay group occurred even though the previously selected wrong answer was visible.

Kulhavy argued that the following three factors were important: (1) the tendency for a test to strengthen responses, (2) subjects forget initial responses following a

delay, and (3) errors interfere with learning correct answers. These three factors, when combined, indicate that the probability of repeating an initial error on a retention test is greater for the immediate feedback groups than for those given a delay prior to feedback. In addition it was found that less time was taken to study feedback presented immediately after answering the item than was taken to study feedback 24 hours following testing. Kulhavy considered this a function of fatigue and frustration. Furthermore, if feedback was reinforcing, one would expect initially correct responses to be repeated. "In fact, the probability of repeating correct responses on the final test was no higher for immediate feedback groups than for the delayed feedback groups". (p. 511) Finally, instructors were given the advice:

> One should take care that learners have thoroughly understood materials before giving them a test. Feedback should be delayed for a day or two, especially if there is an error rate of any magnitude." (p. 511)

Surber and Anderson

In a classroom study using M/C testing and examining the effects of delay of feedback upon retention, Surber and Anderson (1975) detected the importance of delaying feedback in the improvement of scores of high school students. Feedback was in the form of the question and alternatives represented on paper with the correct answer underlined. The 24 hour delay in feedback was contrasted both with immediate feedback following the test and with no feedback. The study

also measured the change from initial wrong answers to correct answers on retention test. Feedback was found superior to no feedback and delay of feedback was found superior to immediate feedback. Verbal ability, as measured by the 36 item French, Ekstrom & Price Test (1963), was found to discriminate between those who effectively used feedback and those who did not; e.g., those who changed previously wrong answers to right answers on retesting.

Surber and Anderson concluded that the delay-retention effect was generalizable to the real world of instruction but applied the following caveat: "It remains to be seen whether the delay-retention effect would appear if a course grade were made contingent upon performance or the materials were made available to students during the retention interval". (p. 172)

Only two studies have been found which openly disputed the DRE phenomenon and attempted to replicate earlier studies for the purpose of indicating that the alleged benefits of delaying feedback were more a matter of chance than the result of an instructional design. The studies, Phye (1970) and Newman, Williams and Hiller (1974), unfortunately seemed to violate several of the essential attributes necessary for DRE. These problems will be discussed in detail beginning with Phye.

Phye

Phye's study (1970), entitled "Verbal retention as a function of the informativeness and delay of informative

feedback: a replication", attempted to re-examine the Sturges (1969) study, which explored the delay retention phenomena as a means for improving long term memory.

Phye's research design was as follows: Eighty-four undergraduate students studied educational psychology in a regular classroom setting. A 30 item, four alternative M/C test was administered and feedback provided in a number of different ways. The 84 Ss were assigned to 18 groups. Four groups received feedback in the form of the question restated. Two of these four groups received feedback immediately after the test while the other two received their feedback following a 48 hour delay. Six groups were used as controls and received no feedback. Of the final eight groups, four groups received feedback in the form of the question stem plus four alternatives (original question) whereas the other four groups received feedback in the form of the question stem plus eight alternatives. Within each of these four groups were the immediate (two groups) and delay (two groups) components. Retention tests were administered immediately following feedback and seven days later. Table 4 summarizes this research design.

Table 4
The Phye (1970) Research Design

# of Groups	Feedback Message	Feedback Timing
2	Stem + 4 answer options	End of test
2	Stem + 4 answer options	48 hours
6	No feedback	N/A
2	Stem + 4 answer options	End of test
2	Stem + 8 answer options	End of test
2	Stem + 4 answer options	48 hours
2	Stem + 8 answer options	48 hours

Feedback in each of the groups was provided by the researcher reading the test items to the subjects as a group and indicating immediately which alternative was correct. The feedback presentation was produced by randomizing the test items and also randomizing the answer options within each item.

Phye provides no information about the sources for the additional four answer options used to compose the eight alternative group; nor the time allowed for feedback, or indeed, the testing procedures used during the initial test or the retention check one week following aural feedback.

Thus, the study differed significantly from that of Sturges (1969) by providing 48 hour delay in lieu of 24 hour delay, group testing and aural feedback in contrast to individualized visual feedback, feedback after the test rather than feedback following each item, and an unspecified method for the development of additional alternatives for feedback.

Phye detected a significant difference ($p < .05$) between those groups receiving immediate feedback and those receiving delayed feedback.

A significant difference ($p < .05$) was also detected between immediate retest scores and seven day retention test scores. The scores ranged from 28.14 - 29.14 out of a possible 30 (ceiling effect) for the immediate retesting group and 24.57 - 27.85 out of 30 for the seven day retention group. A ceiling effect occurred in both instances. The greatest mean

was that of the delay, multiple distractor group.

Thus, although Phye begins with the claim he is providing a replication of Sturges's work, careful reading of both studies indicates a departure from the Sturges design. Having noted the differences from the Sturges design, it is interesting that Phye (1970, p.381) asserts "...certain conclusions drawn by Sturges (1969) apparently need tempering."

Newman, Williams and Hiller

Newman, Williams & Hiller (1974) attempted to produce a definitive study of the delay-retention effect in a totally naturalistic setting. Ninety-four undergraduates enrolled in educational psychology read an assigned article of 3700 words which dealt with a theory about the brain chemistry of short & long term memory. After the alloted 25 minutes of reading time, a 30 item M/C test, composed of 28 four-alternative items and two five-alternative items, was administered for 25 minutes. Four feedback conditions were imposed on the randomly assigned groups (no feedback, immediate feedback, one day delay or seven day delay). Seven days following feedback a retention test composed of 30 randomly ordered items was again administered. No significant differences were detected for any of the feedback conditions, nor was there any significant difference in performance on test items analyzed according

to initial performance or according to item difficulty. A post retention test questionnaire disclosed a tendency for the group receiving immediate feedback to restudy the material but this activity had no differentiating impact on final test scores.

The authors emphasized the desire to maintain external validity, that is, to emulate "real" learning and testing conditions. The subjects studied the material, were informed a retest would occur at a later time, had access to the learning material between tests, and understood their course grade was dependent upon performance. This was one of the few studies not to detect a delay-retention effect or perseveration of error. The reason for the flat performance across treatment groups may arguably have been due to one or more of the following factors:

1. A fixed learning period was used by the students to read the material (25 minutes for all groups).
2. A fixed testing period was used for all groups (25 minutes for all groups).
3. The form of information feedback involved projection of the test item (with the correct answer underlined) on a screen for 15 seconds. Subjects were required to respond via a five button button input box, by pressing the button matching the correct alternative displayed on the screen.

The rationale for requiring subjects to respond overtly to the feedback message was to insure that the feedback was

attended to and processed. It is believed that this procedure partially violated the external validity claim made by the authors and may very well have constituted a relearning situation. Even under the conditions described, feedback in test one did appear to assist students when retested with test two. However, the difference was not significant ($p < .06$).

Summary of Feedback Timing Research

Although Sturges examined the phenomena of feedback and delay using a number of approaches, problems still exist. Most, if not all, of Sturges' learning paradigms were based on learning materials such as definitions and uncommon English words, which were initially presented in a test atmosphere. Immediate retesting, apparently a useful exercise to improve retention, is impractical for most 'academic' evaluation situations. This fact restricts the generalizability and transfer of the findings. In Sturges' early studies, initial presentation, feedback, immediate testing and retention checks were carried out in laboratory learning environments. Presentations were via a Kodak Carousel slide projector and subjects responded on slips of paper. Feedback, practice and retention items were also presented via slides. Thus Sturges' early work possessed the following characteristics:

1. A one exposure learning task that resembled a quiz,
2. Various feedback messages followed immediately by

retesting, a process designed to check specifically on the immediate effects of feedback but which also provided more practice,

3. Retention was measured in a sequenced, precisely timed atmosphere seven days later, and
4. None of the early studies used subject matter for which there was academic credit (motivation).

The 1976 Sturges study was based upon university course material and used a computer managed testing method. Unfortunately, the course instructors were not consistent in the importance they attached to the quiz results, and the time of the retention test varied from one to three weeks after the initial test. This study appears to provide the best indication of approaches to feedback on meaningful subject matter; but, because it is the only computer based experiment so far uncovered, the impact of various feedback forms on learning within CAI environments is still not known.

In addition to the papers of Sturges reviewed in support of DRE, Sassenrath(1968), More(1969), Kulhavy(1972), and Surber and Anderson(1974) also presented findings which supported DRE. These supporting studies indicated DRE has been found to occur under conditions of (1) individual or group testing, (2) tests with multiple choice items, (3) item response and feedback times controlled to periods as short as 10-15 seconds, and (4) using meaningful material drawn from courses or unfamiliar sources. Phye (1970) did

not demonstrate a strong delay-retention effect under conditions of (1) group testing, (2) multiple choice items, (3) unspecificed response times, (4) aural feedback, (4) 48 hour feedback delay, and (5) a sample group that achieved near mastery on the first test. Newman, et al. (1974) failed to trigger DRE when they (1) used new learning material, (2) restricted the study time available to learn the material, (3) fixed the testing time, (4) fixed the feedback presentation time, (5) required an overt response to feedback, (6) allowed access to the material before the retention test, and (7) made the test count as part of the course credit.

Although the majority of the research evidence favours delaying informative feedback, no studies have been found which indicate DRE occurs using material that is part of a university credit course and delivered under CAI conditions -conditions with which the students have become familiar to the point of taking the learning/testing environment much for granted.

This study provided test-item feedback in the form of either immediate feedback following each item, or feedback delayed by 24 hours. The testing and feedback were all within the context of an ongoing 80 hour CAI course in statistics. The study, because of the test content and CAI delivery, is an extension of the work of Sturges (1972, 1974, 1976) as well as the others previously cited who have examined DRE in a classroom setting or used

tests within programmed instructional texts.

The following section reviews research literature reflecting the effect of feedback message design upon long term retention.

C. Research on Feedback Messages

Feedback messages have tended to be terse statements which simply indicated whether the response was right or wrong. For example, Plessey's teaching machine of 1926 was only capable of presenting feedback stating "right" or "wrong". Chemically treated answer sheets which appeared much later, indicated a "Y" or "N" when an answer option was touched by a chemically treated crayon. (Sullivan, Baker, and Schultz, 1967) Early work by Anderson (1967,1971,1972) examined the availability of feedback messages in the context of programmed instruction and later CAI(PLATO). The examples provided within these studies were all of the terse response variety and no guidelines were stated for writing corrective or reinforcing feedback. Author manuals provided to assist instructors in programming CAI courseware on either the PLATO, TICCIT, Philco, DEC, or IBM 1500 systems do not elaborate upon how to write effective feedback messages. Several of these systems have a macro facility which automatically presents feedback messages such as "You are right" or "You are wrong" --aids for easier lesson construction. None of the CAI author support materials available through computer companies discusses programming

to achieve enhanced long term retention.

The first example of feedback message manipulation directed at increasing long term retention was presented by Sturges (1964). In the study the feedback messages were simply the multiple choice questions re-presented with either (1) the correct answer (CA) underlined (the other options remaining) or (2) a cue (CU) which was designed to lead the subject to the correct answer.

The results, using meaningful test material, indicated that under the CA feedback type, those given feedback delayed by 24 hours exibited greater retention than those of the immediate feedback group. If the feedback was of the CU type, however, no significant difference was found between delay modes. In contrasting the feedback types, the group receiving cues achieved significantly greater retention scores after seven days than did those receiving the correct answer. Therefore it was concluded using a cue evokes better retention than simply stating the answer. No differences were found between feedback types or delay modes on nonsense material. Sturges concluded from this study that cues promoted symbolic exploration of alternatives.

In a continuation of the earlier study, Sturges (1969) considered feedback of two types: (a) a stem, plus answer options with the CA underlined, and (b) the stem, with only the CA option underlined. The test base was a 38 item, four distractor M/C test of factual items related to social sciences and two additional questions for samples. The

experimental hypothesis was basically that the delay effect would disappear if the examination of answer options were removed. It was found that removal of this knowledge did remove the effect of delayed feedback upon retention. Increased knowledge does apparently accrue through the examination of incorrect alternatives.

> With the delay of informative feedback subjects appear to respond to more cues, or stimulus aspects of informative feedback, thus learning more about the item, and when these cues can be used in retention, delay improves retention. (Sturges, 1969, p.14)

A question still remained. Did the subjects actually cue from the position and/or number/letter of the distractor or were they actually achieving a deeper understanding of the material?

Sturges (1972a) examined the importance of item construction by (1) either administering or not administering immediate retesting, and (2) delivering feedback with 0 delay, at EOT (end of test), or after a 24H delay.

The four types of feedback were:

1. A replication of the original test item but with CA underlined (RW+),
2. The test item with CA underlined but the distractors randomly rearranged and without letter cues (A,B,C,D) as in the original (RW),
3. The test item formatted exactly the same as its original counterpart but with the CA underlined and the distractors removed (R+), and

4. The item with CA underlined randomly placed without letter cue (R).

Two types of retention, recall and recognition were measured. Recognition scores were derived from a 32 item M/C test which provided a stem (definition) and four uncommon English words as possible matching answers. Selection of the correct alternative was evidence of recognition, whereas providing the appropriate word when answer options were not presented (an open ended question) measured recall. Figure 4 provides examples of feedback messages.

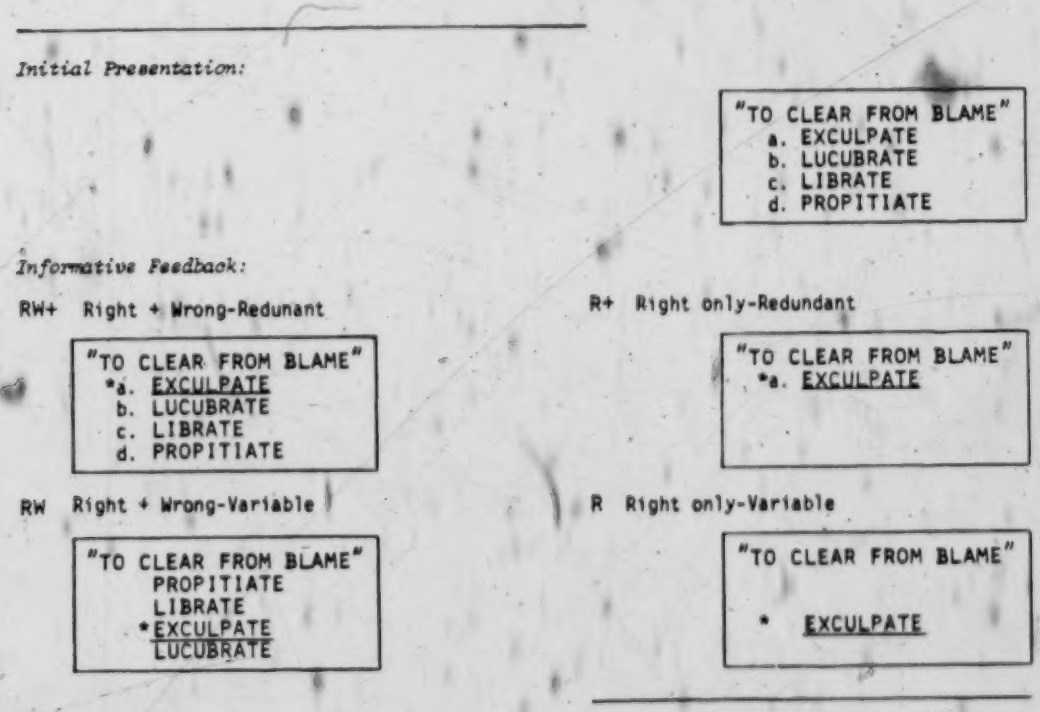


Figure 4: Forms of feedback messages (Sturges 1972, Phase I)

The procedure followed by Sturges (1972) consisted of a testing algorithm with (1) three delay modes, (2) three immediate tests (nil, recall, recognition) for practice and

(3) after seven days a recall/recognition test. Figure 3 illustrates this research design.

The finding regarding immediate retesting following feedback was:

> A significant interaction ($p<.05$) occurred between feedback forms. 'Right+wrong-redundant' appears superior to 'right+wrong-variable' whereas the 'right only-variable' was found superior to 'right only-redundant'.

These results are summarized in Table 5.

Table 5
The Impact of Feedback Messages Immediately after Delivery

Feedback Message		Feedback Message
Right-Wrong Redundant	>	Right-Wrong Variable*
Right Variable	>	Right Redundant*

*Significant at .05 level

The findings on the seven day retention test were as follows:

1. One component of the interaction between form of feedback and type of retention test was significant ($p<.05$); that is, long term recall is best enhanced by using feedback containing all the alternatives whereas long term recognition is best developed by presenting only the correct answer as feedback.
2. In a comparison of feedback types and long term

retention objectives (recall or recognition) it was found that a 24H delay of feedback was superior to EOT delay of feedback for long term recall using "right-wrong variable" in contrast to "right-wrong redundant". Best long term recognition occurred if the delay was 24H and feedback was "right-variable" or secondarily "right+wrong-variable" in contrast to the poorer types - "right-redundant" and "right+wrong-redundant".

On the strength of this study, and if no immediate retesting is possible, long term recall may be best enhanced using 24H delay on feedback of the "right+wrong-redundant", "right-variable" or "right redundant" forms. All feedback types seem equally useful for the content matter under study. For long term recognition 24H delay using either the "right-redundant" or "right-variable" seem equally good. It should be noted that <u>not</u> using immediate retesting seems to cut long term retention as much as 21-49 percent. Table 6 summarizes these findings.

Table 6
Sturges (1972) Phase I Findings

Design Variable	Long Term Objective: Recall	Recognition
Optimal Feedback Message Type:	Right-Wrong Variable	Right Variable
Optimal Feedback Delay Mode:	24 Hour	24 Hour
Optimal Combination of Feedback Delay & Message:	24 Hour + Right-Wrong Variable	24 Hour + Right Variable

In a second phase to this study, Sturges contrasted (1) a feedback message which was composed of randomly ordered answer options and the CA underlined with (2) a representation of the item with only a cue to the answer. (Figure 5 provides examples of these feedback messages.) Different delays, immediate testing, and final retention testing paradigms were the same as in phase one.

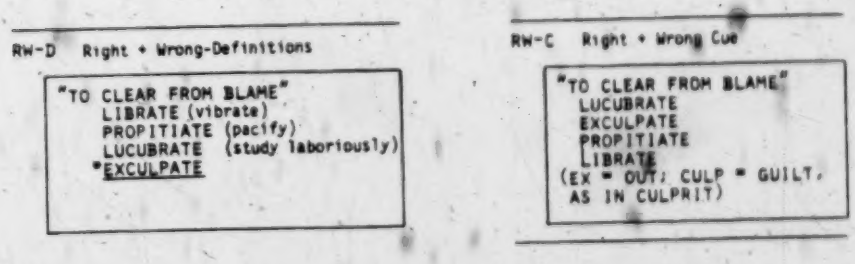


Figure 5. Forms of feedback phase II.

Phase II findings indicated the cue feedback provided better results if there was no immediate test; however, if immediate retesting was employed then "right+wrong-definitions" feedback provided superior results. In addition it appeared that 0 delay in feedback was not appreciably different from that of the other delay modes with cue feedback.

The prescribed practice as a result of these Phase II findings may be described as follows:

1. To obtain the best results on a long term recognition test, delay results by 24H and provide feedback of the "right+wrong-definition" type, then immediately retest using recognition items. If no immediate retesting is possible, change the feedback messages to those of the "right+wrong-cue" type.
2. To obtain best results on long term recall, delay feedback by 24H and provide it in either the "right+wrong-definition" form or "right+wrong-cue" form, then retest either in recall or recognition format. If no immediate testing is possible feedback should apparently be of the "right+wrong-cue" type.

Under zero delay of feedback and immediate retesting it was clear subjects attended best to feedback providing no more than the correct answer. If delay of feedback occurred, it appeared the subjects attended to more cues or stimulus aspects of the feedback, thus learning more about the item.

Delayed cue feedback improved retention. If no immediate feedback test was planned, then superior retention was found only with cue feedback.

These findings indicate that the test designer must first decide upon the retention objective (recall or recognition) and then manipulate feedback to provide the retention desired. If retesting is possible following feedback and optimal recall is desired, it appears subjects can best use feedback that is unambiguous (show only the answer); whereas if optimal recognition (discrimination between alternatives) is desired then feedback which randomizes the answer options is sufficient provided immediate retesting follows. The deeper level of processing which accompanies the cue feedback seems counter-productive if immediate retesting is used.

Sturges concluded that subjects do not acquire much information from any form of feedback if it is immediately presented. It would appear from this study that the presentation of either the CA or WA as feedback does not necessarily amount to useful information. Additional cues are necessary at feedback to improve retention. Further, recall, as well as recognition, increased as a result of a delay of feedback. This improved both the ability to discriminate among distractors as well as provide correct answers from free recall, i.e. minimal cues (the stem).

Thus, in order to improve retention it is necessary that the feedback be of a type that causes the subjects to

infer an associative link between the stem and the answer options, whether right or wrong. In this way the subjects begin to view the alternatives as being organized and positively or negatively related to the stem. One alternative to delaying feedback, might be the use of an immediate feedback message designed to invoke a "novel" mental process to arouse more than a right-wrong concern. Sturges reported that by the use of immediate cue feedback, the results obtained compare favourably with those achieved using a different message form and delaying feedback either until the end of the test or until at least 24 hours had elapsed.

To some extent, less powerful feedback may be enhanced by immediate practice. Yet this may not always be possible or desirable. In the final analysis, the subjects must be lead to re-explore the test material to improve retention.[1] Table 7 outlines the Sturges (1972a) findings

Sturges (1972b), in a continuation of the research on retention improvement through feedback manipulation and delay, compared the use of an instruction to (1) study the correct answer (underlined), (2) study the correct and incorrect alternatives indicated, with (3) the representation of the item with a cue given which would lead to CA selection after some thought (similar to the earlier cue type). It was again theorized that the cue would promote both a careful study of the interrelationships between

---

[1] Personal correspondence with Dr. Sturges.

Table 7
Summary of Sturges (1972) Findings

Testing Procedure	Long Term Objective: Recall	Long Term Objective: Recognition
If immediate retesting follows feedback:		
	Use 24 hour delay of feedback	Use 24 hour delay of feedback
	Correct answer only as feedback message, and	All answer options as feedback message, and
	Immediate test of the recall type	Immediate test of the recognition type
If immediate retesting following feedback is not possible:		
	Use EOT or 24 hour delay of feedback, and	Use EOT delay of feedback, and
	Provide feedback in the form of a cue	Provide feedback in the form of a cue

answer options and a better understanding of the organization of the material would result. A 32 item M/C test was used which was composed of questions in the form of a definition stem and four uncommon English words as answer options. Two testing protocols were used. The first protocol was basically the standard Sturges design which involved administering the test, providing feedback under three delay modes (0, EOT, 24H), immediate retesting and testing seven days later. The second protocol differed from the first only

by providing all of the feedback messages linearly prior to the first test administration. That is, all the feedback messages were presented before the administration of the test. This was done to examine the power of the context (stem with answer options and feedback message) in contrast to seeing only the feedback message. Retesting and seven day retention testing sessions checked first recall then recognition for both protocols.

The seven day retention results reflected the earlier findings; namely, that immediate practice improved the power of some types of feedback but the cue technique achieved best results without the necessity of immediate practice. It also appeared that overall retention was better when immediate practice was of the recognition rather then the recall type.

It was concluded that -

> Instructing the subjects to respond differently at the presentation of informative feedback did affect their retention performance but this effect depended upon the presence of an immediate test as well as the form of the seven day retention test. (Sturges, 1972b, p.4)

> These findings add support to the conclusion that retention of the correct alternatives is facilitated when subjects have had the opportunity to identify relationships among the stem, the correct and incorrect alternatives. (Sturges,1972b,p.5)

Travers, Van Wagenen, Haygood and McCormick

Travers, et al., (1964) designed a classroom based retention study to measure the function of differing task involvement and feedback conditions. The task required students to learn the meaning of German words through either observation or involvement. Each hourly lesson for the first three days of the week concentrated upon the mastery of a 20 word vocabulary. Students who were called upon to answer during the lessons received one of four types of immediate feedback. All were adequate in content but consisted of varying redundancy. The question and answer sessions of the lesson were, therefore, carefully designed. A retention test was given on the Friday following the three days of training.

The study demonstrated a significant difference between feedback groups. "The amount of redundancy is related to the degree to which the task is learned -with greater redundancy favoring learning" (p. 173). Also of interest was the finding that information last transmitted in the feedback messages was best remembered. In addition, "subjects who interacted with the experimenter performed better, not only on the items on which they interacted but also on the items which they learned from observation" (p. 173). It was suggested that participation raises the level of arousal which influences retention of information acquired by observation.

Sassenrath

Julius Sassenrath (1965, 1968, 1969, 1975) examined the effects of different feedback messages and delay of feedback upon retention.

The first study (1965) was designed to determine how retention could best be improved through better feedback message design or delivery following examinations. Two days after a mid-semester 40 item M/C exam, students received one of four types of feedback: (1) no feedback was given --only the total score was provided, (2) examinations were returned with correct answers placed on the blackboard and the class instructed to spend the period checking booklets to detect errors, (3) examination booklets were returned and a discussion took place between instructor and class, and (4) the corrected examinations were returned. Page numbers from a textbook (Cronbach's Educational Psychology) were placed on the blackboard indicating the source of each test item. Rereading these passages independently was a one class period objective.

The feedback groups scored significantly better ($p < .001$) on the end of semester test quiz than did the no feedback group, and the discussion group received significantly more marks ($p < .05$) than did the other feedback groups. Sassenrath concluded students gained the most information from the discussion sessions and therefore were better able to modify previously incorrect thinking.

Sassenrath (1968) again examined both the delay-retention effect as well as feedback message design. In this study, one-half of the immediate informative feedback group received the quiz items again (stem and answer options) with the correct answer underlined. The other half of the immediate feedback group received only the answer options with the correct option underlined. Within each group, one-half were informed that they should try to remember the answer as a retention test would follow. Ten seconds were allowed for Ss to read each item. The same protocol was followed with the delayed feedback group, but 24 hours after the initial quiz. Immediately following feedback, and five days following feedback, the groups were retested using the same items in random order. These retention tests were written at the students' pace.

Findings indicated no significant difference between feedback groups on the initial test, nor the first retention test (immediately after feedback). Yet, after a five day interval, a significant difference ($p < .001$) favored those students receiving the message to learn and to retain materials because of an upcoming future test. They performed significantly better than the group which received only the alternatives with the correct answer underlined.

In an elaboration of the previous study, Sassenrath (1969) examined the effect of providing four types of feedback: (a) item and alternatives with correct answer underlined, (b) alternatives with correct answer underlined,

(c) only the correct answer alternative underlined, and (d) the stem with only the correct answer alternative underlined. Retention was measured both immediately after feedback, and five days later. (The other details of this research design were discussed in part 1 of this chapter.) The findings were: (1) that no significant difference was detected between groups on the initial or immediate retest following feedback, and (2) that subjects who received only the alternatives performed reliably better ($p < .01$) than those who received both stem and alternatives.

Also of interest was the finding that feedback message effectiveness was the reverse of the 1968 study; that is, Ss receiving only the alternatives at feedback performed better than those receiving both stem and alternatives. In terms of the differences between messages, it may be that the stem of the item produces a processing overload, whereas the presentation of the alternatives presents only salient information. These findings were similar to those of Sturges (1969).

The last reported study by Sassenrath (1975) reexamined the data of two earlier studies from the perspective of interference-perseveration hypothesis discussed in part 1 of this chapter (Kulhavy & Anderson, 1972). As a result of the analysis of responses made on initial tests and retention quizzes and thus the replacement of wrong answers by right responses, Sassenrath (1975) concluded from his review:

> If feedback were acting as a reward it might increase the probability that a right response would be repeated, particularly for subjects receiving immediate feedback. The fact that this does not happen while there is a difference favouring delayed feedback after immediate feedback on the R2/W1 [2] and R3/W1 measures is quite conclusive evidence that feedback provides information regarding what is the appropriate response. Thus, it appears that feedback has a greater effect on cognition than incentive motivation in human learning of verbal material. (p. 899)

Phye

In Phye's study (1970), reported earlier, it was noted that feedback messages were either in the form of (1) the stem plus four alternatives (original question) or (2) in the form of the stem plus eight alternatives. Retention tests were administered immediately following feedback and seven days later. Table 4 (referenced earlier) summarizes this research design.

Feedback in each of the groups was provided by the researcher reading the test items to the subjects as a group and immediately indicating which alternative was correct. The feedback presentation was produced by randomizing the test items and also randomizing the answer options within each item.

Phye provides no information about the sources for the additional four answer options used to compose the eight alternative group, the time allowed for feedback, or indeed, the testing procedures used during the initial test or the

---

[2] Where R2 equates to the correct response (R) on the second test (2) and W1 equates to the wrong response (W) on on the first test(1).

retention check one week following aural feedback.

Phye detected no significant difference between groups receiving four answer option feedback and eight answer option feedback, yet did state he achieved, "an increase in performance under delay feedback and an increase in distractor conditions." (Phye, 1977, p. 381)

Kulhavy

In an apparent deviation from the "classical" feedback analyses, Kulhavy and Swenson (1975) experimented with the use of imagery and its effect upon comprehension. In this study, fifth and sixth grade students were requested either to "read carefully" or "form mental pictures" of the instruction, a 20 paragraph text. Tests immediately followed the reading, and one week later the same students were again examined. Learners who used the imagery approach recalled significantly more than the other group. This lead Kulhavy to conclude that if one can supply learners with an efficient memory strategy, it is likely that more will be remembered from the study of text.

Kulhavy, Yekovich and Dyer (1976) examined the relationship of feedback and confidence upon retention using 30 volunteer undergraduate students and a 30 frame programmed instructional course on the human eye. For group 1 feedback was provided by erasing from the answer sheet one of five opaque circles matching each item. A 'T' or 'F' was exposed. Group 2 received no feedback. The time taken for each frame was recorded to the nearest five seconds by the

subject. Retention testing occurred immediately and one week following the initial learning session.

From his earlier work, Kulhavy suggested that (1) feedback operates primarily to correct error responses not to reinforce correct answers (Kulhavy & Anderson, 1972), and (2) little is known about how and when feedback should be used. In a proposed model (see Figure 6) Kulhavy and Anderson declared that confidence is at least as important as the response made in determining the feedback to be provided.

argumentative reaction ensues. A low probability of error repetition is suggested. Those subjects who register either a correct or wrong answer, but have low confidence in the choice are guessing. Therefore Kulhavy, et al. (1976), state:

> If a student is having trouble understanding what he reads, providing feedback after he guesses at an answer should do little to improve comprehension. (p. 524)

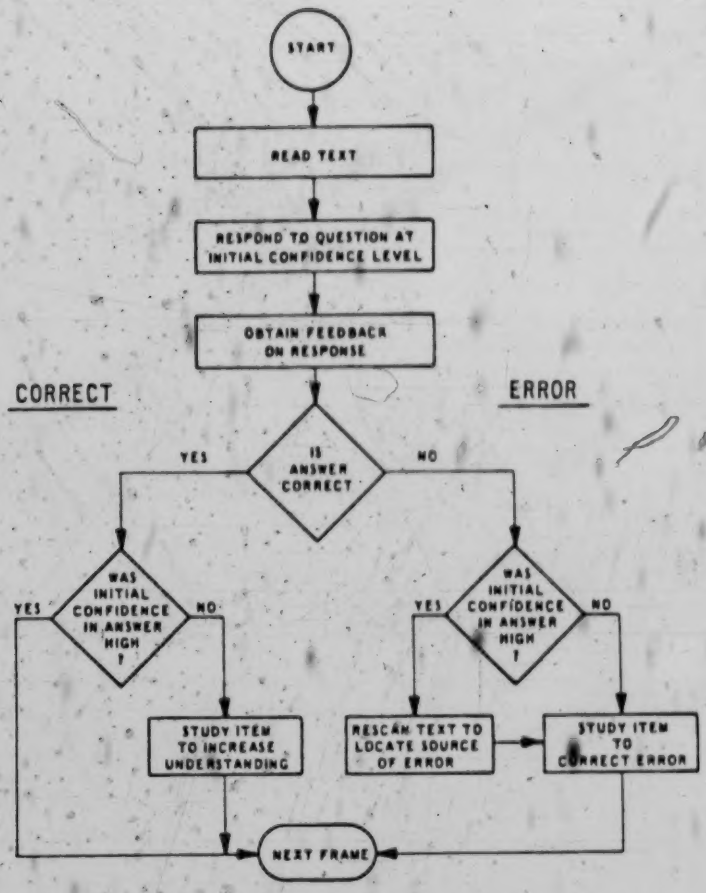


Figure 6. Hypothesized relation between feedback, confidence, and post-response behavior

Those students who are highly confident about the correctness of their choice, and are correct, move on to the next frame without hesitation. Feedback is examined briefly only for direction but not for instruction. However, high confidence associated with a wrong answer results in very careful examination of the question text, and often an

The findings were:

1. Subjects receiving feedback are more likely to repeat right answers ($p < .05$).
2. On a delayed test, subjects tested immediately after the learning session scored higher ($p < .001$) than those not immediately tested. (also Sturges, 1972a)
3. Of those students receiving feedback, the high and the low confidence persons who made correct responses remembered those correct answers best ($p < .05$). This was an unexpected finding for those with low confidence.
4. Groups receiving feedback were found more likely to correct an error on immediate and delayed tests ($p < .05$). "When a subject receives feedback following a high confidence error, he shows a marked tendency to be able to correct himself on an immediate test, and to a less systematic degree on a delayed measure". (p. 526)
5. Subjects spent more time on feedback in error conditions than when correct ($p < .01$). As confidence ratings increased, the time taken to examine feedback following errors increased and less time was taken to examine feedback following correct responses. ($p < .01$)

The researchers concluded that feedback value is directly related to the learner's confidence in his answer. If a high confidence error occurs, more time is spent reprocessing the available information. The longer people study the correct information, the longer they remember it. Thus, subjects are most likely to study an item longer when a confident response turns out to be wrong.

The implications for instructional design are that content must be appropriate to the student and must be comprehended prior to testing since feedback is only valuable as a corrective mechanism when the student understands what was read. It is noted:

> We feel that these findings have applicability for the more sophisticated instructional delivery systems -- especially those involving computer control. Varying feedback procedures and content on a frame-by-frame basis could yield substantial gains in the amount the subject is able to learn from the lesson. (p. 528)

Kulhavy and associates derived their suggestions from research findings that used, what would be considered, crude measurement conditions for CAI instruction. Thus it seems reasonable to suggest that further research be undertaken using computer controlled instructional delivery systems.

Summary of Feedback Message Research

The following is a summary of the feedback research literature reviewed in this proposal. It is based, in part, upon Kulhavy's (1977) recent interesting paper, which provides an integration of the work done on feedback processes, especially as these results apply to written lessons and the design of instructional materials.

A review of feedback research must canvas issues; such as, whether reinforcement occurs in a behavioral sense, the availability of feedback, and learning from feedback.

Reinforcement was examined by Anderson, et al., (1971) They discovered that during a lesson, students seemed to

learn more from feedback provided following an error rather than feedback provided following a correct answer. This is a reverse of what a reinforcement theorist would expect. In addition, delay-retention effect (DRE) studies concluded that immediate feedback did not produce the desired degree of retention when compared with a delay of 24 hours (Kulhavy and Anderson, 1972). Researchers who advocated intermittent reinforcement schedules have found from the outset that learners were clearly at a disadvantage (Anderson, 1967). Experiments involving extrinsic rewards, such as payment for better academic performance, have also failed to produce improved results (Sullivan, Baker & Schutz, 1967). Kulhavy (1977) concluded it is difficult to find data indicating that feedback, following written instruction, functions in the manner predicted by Skinner and others.

On the question of availability of feedback, evidence is clearly against allowing the learners to see feedback before responding (Anderson, 1972). It is also against designing questions which are so heavily cued or prompted that students do not have to attend carefully to the text (Anderson, 1967). Students who copy the answers take less time, commit fewer errors but, unfortunately, retain less (Anderson, 1971).

Where feedback has been found to increase learning, the intellectual operation differs, depending on whether the submitted response is right or wrong. Provided the correct response occurs as a result of something other than a random

guess, it is clear that it will perseverate over a number of tests (Kulhavy & Parsons, 1972 and Kulhavy & Anderson, 1972). It is assumed that feedback under this condition confirms overall comprehension rather than to clarify a specific term.

Kulhavy (1977, p.221) stated that "one of the reasons why corrective aspects of feedback have received so little attention is simply that many studies fail to analyze errors and correct responses separately". Travers, et al., (1964) and Anderson, et al., (1971) indicated the benefit of first learning that the response was wrong and then learning the correct response. Delay is needed to eliminate the old concept and overlay the new one. DRE apparently enhances the impact of corrective feedback since it reduces the memory of initial error responses. Proactive interference, if present, blocks or obfuscates correct answer acquisition under immediate feedback conditions. Evidence was found to indicate error perseveration decreased if the learner was allowed time to forget his initial wrong reponse (Kulhavy & Anderson, 1972; More, 1969; Brackbill, Wagner & Wilson, 1964; Sturges 1969, 1972; Sassenrath, 1975). List learning experiments involving proactive interference provided similar results (Underwood & Freund, 1968).

The question of comprehension enters into the DRE equation, however. Although not explored in detail, the findings of Kulhavy & Parsons (1972) and Kulhavy (1977, p.223) suggested "feedback will have only a minimal effect

if the learner is unable to comprehend the instruction or fit it into some existing information framework".

The conclusions reached by Kulhavy were: (1) entry skills of learners must be sufficiently high to make profitable use of instruction and feedback; (2) design of instruction, in particular the questions, must ensure appropriate levels of cognitive processing (i.e. no over prompting or copying); and (3) feedback must have both informative and corrective power such that the learner recognizes an error and engages in a remedial process. Kulhavy (1977, p.225) observed, "because computerized instruction allows such a wide range of strategies for each response, the question of how one most effectively matches feedback parameters with response characteristics is indeed an important one".

It is because most studies have examined feedback and retention, generally in binary terms, i.e., the response was right or wrong, that little was learned about how the question was perceived and the probabilities associated with repeated similar responses. Kulhavy (1977) argued for more sophistication in the test taker model since it appears that question answering begins with an assessment of potential answers and the assignment of a hierarchy of confidence. The final selection of a probable right answer is made, provided the answer is not obvious, from the context of the question, the content of the stem, the availability of the answer, or the selection of alternatives.

> The case for studying how learner characteristics ... influence the use of feedback seems well made. This appears to be a prime area for future research, one which may shed new light on an old, and well-turned field of instructional psychology. (p. 229)

The next section discusses the research design and subjects used in this study.

IV. Research Methodology

Introduction

The following sections describe: the research questions, the design of the study, the instructions to subjects, and the sample. Within the section on the design of the study are described the STAT1 computer assisted instruction course, the learner characteristics, and the instrument used to explore the effects of feedback delay and feedback message design within the context of computer-assisted instruction. This section also indicates: the characteristics of learners who took the STAT1 course on the IBM 1500 system, the STAT1 author support programmes which provide the means for tracking all students' responses throughout the course, and the STAT1 exam developed to evaluate learning on the 't' test segment.

A. Research Questions

This study examined two commonly used instructional design constructs specifically for their effects on long term retention. The questions asked were:

1. Does immediate feedback result in better long term retention than feedback delayed 24 hours?
2. Does a feedback message which consists of underlining the correct answer to a multiple choice question result in better long term retention than a feedback message which is designed to be a cue to the correct answer?

In addition to the delivery of instruction, the IBM 1500 system has an important research role as a data collection device. These data were used to satisfy additional instructional questions.

The supplementary questions were asked to determine if the two key variables under study, i.e. feedback timing and feedback messages, have an effect upon:

a. the mean confidence students assign to their responses,

b. the mean latency time subjects require to produce responses, and

c. the mean latency time taken to read a feedback message.

B. Design of the Study

The Computer Assisted Course: STAT1

The computer assisted instructional course, STAT1, is a basic statistics course designed to prepare graduate students in education to handle various research problems typically encountered in education. The course, although it initially appears linear in progression to the student, is essentially under learner control. This means that by using special keyboard operations, STAT1 students may move at will within each segment (chapter) of the course or in fact transfer between segments in the course. Thus, students may determine the order with which they will study the chapters

and progress through the course or review as desired. The authors have designed decision points to encourage students to reexamine previous materials or to study prior to examination. To facilitate student movement within the course, information sheets are provided which detail all segments and their internal headings.

The course, composed of approximately 100,000 computer instructions, required about 3,000 hours of time to design, programme, and revise over a period of four years. Student terminal time to complete the course ranges from 29 to 160 hours; the average completion time is 69 hours. Approximately 3,000 responses per student are registered during the course. In addition to covering the course material, students must also complete 11 tests interspersed between selected course segments; seven of these tests are administered in CAI mode. The total terminal time taken to write these seven exams ranges from 2.9 to 21.9 hours, for an average of 8.3 hours. Under CAI mode, students may sign-on to take an exam at any point in the course. The score obtained will be the mark accumulated on the first pass through the exam. Generally students take the examinations following the related course segment. The examination is created by randomly ordering all the items in a fixed pool. As a result, every student receives an exam with items uniquely ordered. Although it is possible for students to sign-off during an examination, the start point upon sign-on will be at the last unattempted item. No review of

course material or movement out of the exam is possible once an exam is begun. The feedback design employed throughout the STAT1 exams provide the student with a terse message (stating the correct answer) immediately following the student's response to a question.

The STAT1 instructor support programs provide the following information:

a. a re-creation of all terminal screen displays as seen by the student,

b. the anticipated answers to each question, as programmed by the course author, and an indication as to whether the anticipated answers are correct or incorrect,

c. the total number of student responses accepted as correct, wrong, or unanticipated, and

d. opposite each response category, is displayed the student's ID along with a number indicating if this is the student's first, second, or Nth attempt at the question.

Hunka, Romaniuk and Maguire (1976) indicated students not only readily accomplished the educational goals of this basic statistics course, but that they also saved themselves and their instructors approximately 24 hours of lecture time as well as 84 hours in laboratory sessions. Furthermore, subjects indicated a high degree of satisfaction with the course, and the instructors expressed pleasure at being able to assist individual students at the terminal or to mark

and discuss lab assignments on a one to one basis. In general, therefore, more personal contact was assured during the periods when students could benefit most.

In summary, STAT1 represents one of the few comprehensive courses that establish the viability of CAI instruction. The length, instructional design, complexity of subject matter, and demonstrated success place it in an area with few peers (Kearsley, 1976). Learner control and an apparent ability for the courseware to 'stand alone' support the wisdom to invest several thousand hours in its creation and continued optimization.

The Instrument

The instrument used to explore the effects of feedback timing and message modification was a twenty-three item M/C test on the topic of 't' tests. This test has demonstrated consistent levels of difficulty over the past three years. The mean test score is approximately 15 correct out of a possible twenty-three. The standard deviation is 2.4.

This test possessed the following characteristics:

1. Twenty of the twenty-three test items were presented in a random order to each student.
2. All items were in an M/C format with four answer options (a,b,c,d).
3. Two feedback messages were constructed for each item. One was a re-presentation of the question with the correct answer underlined, and the other was a cue designed to lead the subject to understand the question

and to recognize the correct answer option. The cues were written by the researcher and subjected to scrutiny by the three authors of the STAT1 course for amendment or approval.

4. Feedback was either provided immediately following the student's response to a question or was stored for presentation 24 hours later. If the subject was in the immediate feedback group, the total test score was provided after feedback on the twenty-third item, otherwise the total test score was saved along with the feedback messages for delivery the next day.
5. Provision was made to ellicit the confidence that each subject had in the 'correct' response to each question, as well as the confidence the subjects had that each of the other three answer options were incorrect. A seven point continuum (1-7) was presented. By pointing to 7 the subject indicated absolute certainty that the answer option presented was correct or wrong. By pointing to 1 the subject indicated he was not certain if the answer option presented was right or wrong. Appendix 1 contains a sample question accompanied by sample confidence measures.

A second retest was administered a week following the first test in an attempt to check the long term retention of the subjects. The characteristics of the second test were:

1. The first twenty test items were randomly ordered for presentation,

2. The student's confidence regarding the four answer options was measured following each question completed by the subject. (As described above.)
3. Feedback was delayed until after the last item was answered and was then presented in the form of the correct answer option underlined. After the feedback message for the last question was displayed the total test score was presented.
4. The last three questions in the exam, which required calculations to be performed based upon a specific formula, were changed by altering the raw data presented with the questions.

In summary, the instruments were used under the following conditions:

1. All students had a portion of their final course credit dependent upon their success on these exams. All computerized end of segment STAT1 tests contributed a possible 6% toward the final grade. In this particular case, the first 't' test exam carried a weight of 4% of the final mark and the second (retention) test carried a weight of 2% of the final mark.
2. Both exams were administered on the IBM 1500 system. Students controlled the time taken to respond to each question and to study the feedback.
3. During the first test administration, students in the immediate feedback section received their total test score immediately following completion of the test.

Those in the delayed feedback group received their total test scores after presentation of all the feedback (24 hours later).

4. Confidence measures were made of answer options selected by the student and of each of the other answer options not selected.
5. The IBM 1500 system recorded the time each student spent prior to responding to each question and also the time taken to read feedback messages.
6. Finally, comprehension, as well as recognition measures, were obtained. This was made possible by changing the raw data required for the solution of three items on the long term retention test. The algorithms underlying the solutions to these questions were left intact.

The procedures followed in this study differed significantly from previous research activity in this field in the following ways:

1. The testing technique was integrated into an existing, stable computer assisted instruction system. The teaching behavior of instructors was not challenged or subjected to change. They were given complete freedom to interact with subjects as they would in any normal CAI environment.
2. Students were not restricted in the time they spent attending to the test items or the feedback provided. In fact these times were collected by the computer and used for analysis.

3. The two exams were administered under the same CAI conditions.
4. A standardized time interval (one week) existed between the first and second testing sessions.
5. The final test score was provided irrespective of group membership.
6. Confidence measures were made of the distractor selected as correct and the other distractors not selected as correct.
7. The test was based upon a credit course within a regular CAI learning environment.
8. The test scores formed a portion of the final course mark.
9. Both recall and recognition of material was tested.

When the students took the first exam, they were assigned to one of eight treatment groups depending upon the student ID numbers. These numbers were assigned to them for use on the IBM 1500 system. As a result of the modulus 8 value of their ID number they received one of two possible treatments in each of the following three categories:

1. student confidence was either solicited or not solicited with respect to the four distractors for each question,
2. feedback was either provided prior to the next question or was presented the next day, and
3. feedback was either in the form of correct answers (CA) underlined or as a cue (CU) to the correct answer. An example of a test item, question 2 from the test, is found in

Appendix A. The matrix of the design used to assign students to the eight treatment conditions is presented in Table 8. (Appendix C contains a schematic of the research design.)

The Sample

This section describes both the sample and the behavior of the sample during the research process.

Subjects enrolled in Educational Psychology 502 and Educational Administration 511/512 during the 1978 Special Session at the University of Alberta, Faculty of Education, were randomly assigned to one of the eight cells in the research design (Table 8). By the end of the Special Session a total of 60 students had completed the tests. Unfortunately, several problems related to data acquisition and student drop-out reduced the number of subjects to 50.

Table 8
The Four Factor Experimental Design

	Feedback Group				Total
	Immediate		24 Hour Delay		
	CA[3]	CU	CA	CU	
Confidence	6	6	5	9	26
No Confidence	6	8	5	5	24
Total	12	14	10	14	50

[3] CA=Correct Answer, CU=Cue

The CAI environment differed from the usual classroom environment. In the CAI environment some students may be progressing through tutorial portions of the course while others are either doing their lab exercises or taking a test. In addition, because of the self-paced nature of CAI, the tests under investigation in this study were administered over a four week period during each of the two six week Spring and Summer Sessions. The examinations were 'open book', which meant that during the exam students could consult any notes and textbooks at their disposal. Note taking was also permitted during the exams. No examples of cheating were discovered despite this very liberal approach. Students were also able to question the instructors about problems or concerns encountered during the exams. Several students were provided with very thorough explanations of the questions and their solutions. As an example, two students in the cue group were told which questions they answered correctly (the cue feedback did not tell the student if his response was correct). Additional information was provided to the student only if requested. Finally, no review of the first exam was permitted by the IBM 1500 system before the second (retention) test was taken.

An example of a test item, confidence measures, and feedback is found in Appendix A.

Instructions to Subjects

At the commencement of the test, the subjects in the confidence group were informed of the type of feedback to be received and given directions on how to indicate their confidence levels. They were not informed of the different treatment groups in existence. Those not in the confidence group were informed of the type of feedback to be received. They also were not told of the other treatment groups.

Those in the 24 hour group were requested to return one day later to review the exam and to receive feedback. All subjects were also advised that a retention test would be given one week following feedback, since through self-pacing all STAT1 students would eventually come to know of the second test.

Note taking during the tests was not controlled. However, at the beginning of the retention test, subjects were informed that notes from the first test were not to be used. From observation, it appeared that students rarely took the time to write out each question in detail.

C. Analysis of Data

The following data were obtained:

1. Individual test scores- an item by item record of student performance at the end of chapter test and retention test.

2. Response latency- the time in tenths of seconds required by subjects to respond to each of the test items at test and retest.
3. Confidence measures- the confidence subjects expressed (on a continuum from 1 to 7) in their selected answer as well as on the remaining three answer options not considered correct. These data were available for one-half of the sample at test and retest.
4. Feedback latency- the time in tenths of seconds that each feedback message was displayed on the computer terminal during the test and retest.

Feedback Timing Analysis

Question 1:

1. Does immediate feedback result in better long term retention than feedback delayed 24 hours?

The test score means for the entire sample are summarized in Table 9. While both feedback groups improved appreciably on the retest, no clear pattern emerged favouring any one feedback group. The statistical analysis which follows examines the test scores in detail. (Figure 7)

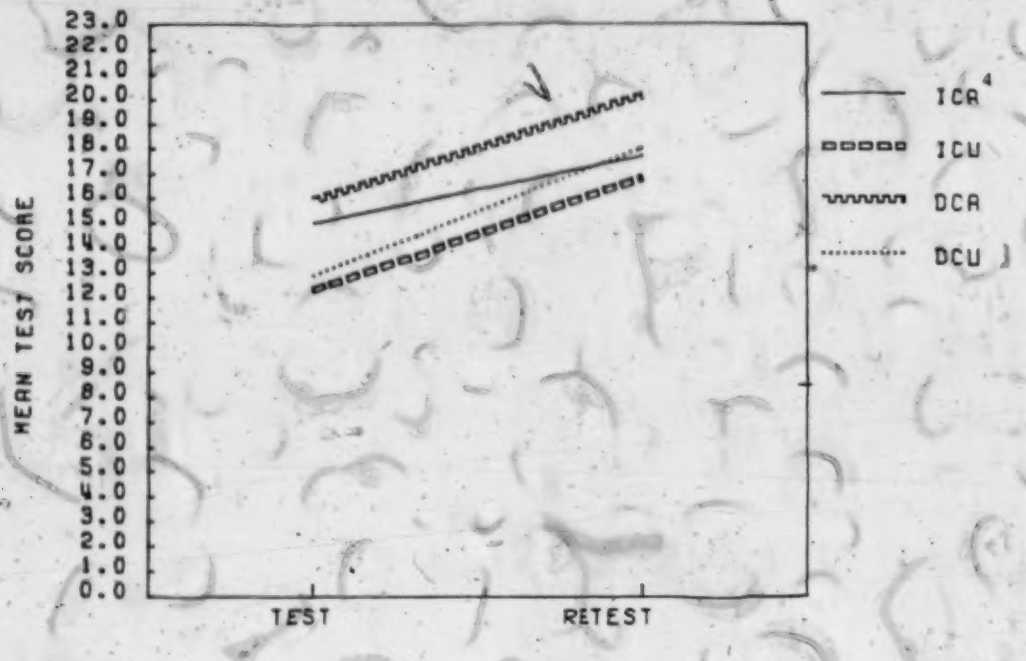


Figure 7. Mean test scores of treatment group by time of testing

---

[4] ICA=immediate correct answer, ICU=immediate cue, DCA=delayed correct answer, DCU=delayed cue.

Table 9
Test and Retest Score Means

Confidence Measures		TEST MEAN	TEST SD	RETEST MEAN	RETEST SD
Immediate	Correct Answer	13.5	3.89	18.5	2.88
	Cue	13.8	2.64	16.3	2.58
24 hr Delay	Correct Answer	16.6	4.34	20.4	2.79
	Cue	12.5	3.78	17.1	3.30
No Confidence Measures		**TEST MEAN**	**TEST SD**	**RETEST MEAN**	**RETEST SD**
Immediate	Correct Answer	15.3	1.03	17.2	2.64
	Cue	11.8	4.09	17.7	2.25
24 hr Delay	Correct Answer	15.2	4.21	19.8	2.39
	Cue	14.6	2.88	18.8	4.09

A four-way analysis of variance on test scores, with repeated measures on Factor D (test,retest), was performed using 40 of the 50 available subjects. This equal 'n' analysis was achieved by randomly dropping subjects from cells containing more than five subjects. Table 10 presents these test score means.

Table 11 indicates a significant difference existed between the two treatment groups on Factor A (correct answer, cue) and factor D(test,retest). Since the analysis confirmed that the questions regarding confidence in answers had little effect upon test scores, the eight cell matrix was collapsed to four

cells by merging the confidence factor. These merged scores are presented in Table 12. On the subsequent three-way analysis of variance, with Factor C repeated, a significant difference was found on Factor B (correct answer, cue) and Factor C (test, retest) The results are found in Table 13.

Table 10
Equal 'n' Test and Retest Score Means

Confidence Measures		TEST MEAN	TEST SD	RETEST MEAN	RETEST SD
Immediate	Correct Answer	14.6	3.13	18.2	3.11
	Cue	14.2	2.77	16.0	2.74
24 hr Delay	Correct Answer	16.6	4.34	20.4	2.79
	Cue	11.2	2.17	17.2	3.19
No Confidence Measures		**TEST MEAN**	**TEST SD**	**RETEST MEAN**	**RETEST SD**
Immediate	Correct Answer	15.4	1.14	17.6	2.70
	Cue	10.4	3.71	17.6	2.70
24 hr Delay	Correct Answer	15.2	4.21	19.8	2.39
	Cue	14.6	2.88	18.8	4.09

Table 11
Summary of Analysis of Test Scores

Source of Variation	df	MS	F
Between Subjects	39		
A (CA, Cue)	1	99.01	7.185*
B (Immediate, Delay)	1	30.01	2.178
C (Confidence, No Confidence)	1	.31	.022
AB	1	2.11	.153
AC	1	6.61	.480
BC	1	7.81	.567
ABC	1	27.61	2.003
Subjects within groups	32	13.78	
Within Subjects	40		
D (Test, Retest)	1	348.61	66.92**
AD	1	7.81	1.50
BD	1	4.51	.87
CD	1	2.81	.54
ABD	1	.61	.12
ACD	1	5.51	1.06
BCD	1	7.81	1.50
ABCD	1	27.61	5.30*
Time X subjects within groups	32	5.21	

*F.95 (1, 32) 4.15
**F.99 (1, 32) 7.50

Table 12
Merged Test and Retest Score Means

		TEST MEAN	TEST SD	RETEST MEAN	RETEST SD
Immediate	Correct Answer	15.0	2.26	17.9	2.77
	Cue	12.3	3.68	16.8	2.70
24 hr Delay	Correct Answer	15.9	4.09	20.1	2.47
	Cue	12.9	3.00	18.0	3.56

Table 13
Summary of Analysis of Mean Test Scores

Source of Variation	df	MS	F
Between Subjects	39		
A (Immediate, Delay)	1	30.01	2.24
B (Correct Answer, Cue)	1	99.01	7.37*
AB	1	2.11	.16
Subjects within groups	36	13.43	
Within Subjects	40		
C (Test, Retest)	1	348.61	57.06**
AC	1	4.51	.74
BC	1	7.81	1.28
ABC	1	.61	.10
Time X subjects within groups	36	6.11	

*F.95 (1, 36) 4.11
**F.99 (1, 36) 7.39

Post hoc analyses, using Scheffe's method,[5] indicated the immediate feedback group and the delay feedback group were not different, based on the mean scores on the end-of-chapter test ($p < .45$). However, on the retention test, the delay feedback group had a higher mean score than the immediate feedback group ($p < .09$). This lends some support for the position that delaying feedback may be of benefit. The findings in this study indicate the difference between mean scores for the immediate feedback groups and delayed feedback groups was not significant.

---

[5] Winer, 1971, pp.563-567. The significant level used for all comparisons was $p < .05$.

Questions Modified on the Retention Test

Questions 19, 20, and 21 were modified for the retention test. This was done by changing the data necessary for their solutions while leaving the algorithms for their solutions intact. In addition to these answers being different on the retention test, the answer options (a,b,c, or d) were also changed. These changes were made to determine if the subjects were learning the answers by rote or were achieving a deeper understanding of the test material. The test score means for questions 19, 20, and 21 for all groups are summarized in Table 14.

Table 14
Mean Test Scores on Questions 19, 20, & 21
for All Treatment Groups

Group	Mean Score* Test	Retest	n
ICA	2.2	2.3	5
ICU	1.6	1.6	5
DCA	2.0	2.4	5
DCU	1.6	2.5	5

* maximum score = 3

A three-way analysis of variance, with Factor C (test,retest) repeated, was performed on the mean test scores for questions 19, 20, and 21. The results, summarized in Table 15, indicated a significant difference between the means on Factor C (test,retest).

An examination of the subjects' responses disclosed no evidence of rote learning. Few examples were found of subjects who provided the same responses on both the test and retest. Hence, subjects recognized the difference between the questions. The findings indicate that test scores on questions 19-21 increased generally. No group derived a significant advantage from the treatment received.

Table 15
Summary of Analysis of Mean Test Scores for Questions 19, 20, & 21

Source of Variation	df	MS	F
Between Subjects	39		
A (Immediate, Delay)	1	.80	.73
B (Correct Answer, Cue)	1	3.20	2.90
AB	1	1.25	1.13
Subjects within groups	36	1.10	
Within Subjects	40		
C (Test, Retest)	1	2.45	5.48*
AC	1	1.80	4.02
BC	1	.20	.45
ABC	1	.45	1.01
Time X subjects within groups	36	.45	

*F.95 (1, 36) 4.11
**F.99 (1, 36) 7.39

Discussion: Delay-Retention Effect

In the review of literature related to the delay-retention effect, it was reported by many researchers that a delay in the feedback message will produce (1) corrective action (to eliminate errors) and (2) reinforce correct responses. Sturges indicated that the "more complex the task the greater the superiority of retention with delayed informative feedback" (Sturges, 1969, p.14). Sassenrath concluded there was "mounting evidence that delayed informative feedback does not retard learning and may enhance delayed retention" (Sassenrath, 1968, p.72) These findings, first discussed by Brackbill (1962), have been confirmed by others, (Anderson, 1971, 1972; More, 1969; and Kulhavy, 1972).

Kulhavy suggested the reason immediate feedback did not apparently lead to better long term test results (retention) was because fatigue and frustration exist during the testing session with the result that the learner does not process the presented data (feedback message) as carefully as he should. In the case of an error, he proposed that the erroneous relationship (A-B) was not effectively converted to another relationship (A-C) because of proactive interference. A time delay followed by feedback, it was suggested, produced superior results because the interference, fatigue and frustration had time to dissipate. (Kulhavy and

Anderson, 1972)

The findings of this study do not confirm or deny the existence of the delay retention effect. The change in student test scores appears to be in the direction of supporting the benefits of delaying feedback. The reasons for this may due to some of the following:

a. The learning environment and student behavior was highly stressful for students during this investigation. In a period of six weeks the graduate students, many of whom appeared to possess weak mathematical skills, were attempting to complete a six credit laboratory course in a critical core subject.

b. Many students described periods of insomnia, fatigue, depression, and going 'blank'.

c. A prevailing concern was the impact of STAT1 scores upon the final grade.

These effects appeared to produce an unusual determination to succeed. As a result, students worked extremely hard at mastering the material prior to the testing sessions. Another factor may have been the personal sessions with the instructors. This form of feedback may have neutralized some of the impact believed to occur with delay. The routine followed within this CAI environment was to provide as much assistance as the student requested. An additional problem arose from the initial differences

between treatment groups on the first test, a condition which made subsequent comparisons difficult. The small sample size was also an unfortunate circumstance.

This researcher concluded the delay retention effect has potential but its effect is apparently less pronounced in a learning environment which (1) allows for an on-demand, detailed one-to-one feedback with the instructor, and (2) is composed of high achieving, highly motivated students. One thing seems certain, the delay design did not retard learning. On this point it is of interest to recall Sassenrath's comment that:

> Although the difference in retention is not large in the absolute amount the psychological importance of the difference, contrary to the accepted principle that immediate reinforcement as opposed to delayed reinforcement, produces superior results and therefore superior retention. (Sassenrath, 1969, p.176)

This study did not produce the differences in test scores that other researchers would have predicted, but it did indicate an immediate reinforcement schedule is not an instructional necessity.

The study has also provided some answers to the questions posed by Surber and Anderson (1974):

> It remains to be seen whether the delay retention effect would appear if a course grade were made contingent upon performance, or materials were made available to students during the retention interval. (Surber and Anderson, 1974, p.172)

This study left unrestrained the routine instructor and student activities and interactions, and access to learning

materials and notes taken during the exam. As a result, the delay retention effect may have been dampened. In spite of these mixed findings, it is believed that CAI instructional designers may have to reexamine the principle of immediate feedback during examinations. In CAI, immediate feedback on examinations has almost become a universally applied treatment. Consideration should be given to other forms of feedback algorithms that could be possible in light of the subject matter, students, and learning environment. The delay retention effect may be extremely useful in circumstances other than the one existing in this study.

Feedback Message Analysis

Question 2

2. Does a feedback message which consists of underlining the correct answer to a multiple choice question result in better long term retention than a feedback message which is designed to be a cue to the correct answer?

The mean test scores used in the question 1 analyses were re-examined for the question 2 analyses. As indicated (Table 11), a four-way analysis of variance detected no difference between groups except on Factor A (correct answer, cue), and Factor D (test, retest). Since the analysis confirmed confidence measures had no measureable effect upon the test scores, the eight cell matrix was collapsed to four cells by merging the confidence factor.

On a three-way analysis of variance, with Factor C (test, retest) repeated, it was again confirmed that the

correct answer (CA) group was significantly different from the cue(CU) group (Factor B), and that a significant increase in scores occurred at retest (Factor C). Table 13 summarizes these findings. An analysis of means on Factor B (feedback types), using Scheffe's method, indicated the correct answer group was significantly different from the cue group ($p < .005$) based on the end-of-chapter test scores. However, on the retention test the difference in mean test scores was not significant ($p < .11$).

A Scheffe's comparison on Factor C (test, retest) indicated a significant time effect existed for all treatment groups. The differences between the end-of-chapter test scores and the retention test scores for all groups was significant at the .0001 level. All groups performed significantly better on the retention test than on the end-of-chapter test.

Discussion: CA and Cue Feedback Messages

The review of literature indicated that a cue feedback message would result in greater processing and longer retention than a feedback message consisting of an underlined correct answer to a multiple choice question. This finding was first advanced by Sturges (1972). A theory describing how data are processed and retrieved proposed by Norman and Bobrow (1976, 1977), Norman and Rumelhart (1976), and Williams (1978), provides support for Sturges' findings. The theory suggests that a cue, if attended to and processed, causes the subject to engage in multiple

retrievals and to perform conscious or unconscious comparisons among competing answer options. An effective cue, it can be argued, results in a better understanding of the interrelationships among answer options and with the question stem. It was postulated that the results of this mental processing lead the student to select the one answer option with the highest probability of being correct. This study indicates that an effective cue may be difficult to compose. In this investigation some students expressed confusion and frustration with cues while others used them without comment. It appeared that students who understood the material infrequently commented about a cue in contrast to those who had a lesser understanding of the subject. The use of cues was an entirely new experience for students. Earlier feedback forms only stated, "Right" or "Wrong". It would seem, because of the many dimensions which exist in problem solving, effective cues must be carefully tailored to suit the student. For example, one of the statistics instructors who has many years of experience teaching graduate level students, reported that students who do not understand an algebraic proof will often understand a geometric proof. As a general rule, it is common to try different teaching strategies to bring students up to a mastery level on the material. The need to have different cues may be considered a sub-set of this general approach. In the future it may be possible to model the learner with sufficient accuracy to predict which type of a cue will be

most effective. At present, more needs to be known about cue construction before the concept can be either fully accepted or rejected as an instructional approach. The findings in this study do not support or reject the practice of writing cues or simply providing a message with the correct answer underlined.

Additional Analyses

In this study two variables have been under examination: (1) feedback timing (immediate delivery, 24 hour delay), and (2) feedback messages (an underlined correct answer, a cue to the correct answer). With respect to these two variables, feedback timing and feedback messages, supplementary questions were asked in order to determine if these two variables have an effect upon:

a. the mean confidence students assign to their responses,
b. the mean latency time subjects require to produce responses, and
c. the mean latency time taken to read a feedback message.

Confidence Values

This study solicited confidence measures from one half of the subjects taking the examination on 't' tests (N=20). The procedure (as described earier) required these subject to provide an answer to the test question and, following that action, to indicate the confidence placed in the correctness of the answer by pointing to a position on a continuum from 1 (not certain) to 7 (absolutely certain). The computer stored the student's answer and the confidence value indicated. The student then either received feedback on the answer or received no feedback, and proceed to the next question. The mean confidence values are presented in Table 16.

Table 16
Mean Confidence Values for All Responses

Group	Test	Retest	n
ICA	5.80	6.13	115
ICU	5.02	5.75	115
DCA	5.48	5.91	115
DCU	5.88	5.81	115

n = 115 = 23 items/test X 5 subjects/group

In a three-way analysis of variance, with Factor C (test,retest) repeated, a significant result was found for Factor C. (Table 17)

Post hoc analyses indicated that in the seven day interval between tests, the immediate feedback group significantly increased in confidence ($p < .02$).

Table 17
Analysis of Mean Confidence Values

SUMMARY OF ANALYSIS OF VARIANCE SOURCE	SS	DF	MS	F	P
BET SUBJ	23.422	19			
A (I,D)	0.085	1	0.085	0.06	0.805
B (CA,CU)	0.460	1	0.460	0.34	0.567
AB	1.336	1	1.336	0.99	0.334
SUBJ W GROUP	21.541	16	1.346		
WITHIN SUBJ	5.160	20			
C (T,RET)	1.257	1	1.257	6.52	0.021
AC	0.308	1	0.308	1.60	0.224
BC	0.0054	1	0.0054	0.03	0.870
ABC	0.505	1	0.505	2.62	0.125
C X SUBJ W G	3.085	16	0.193		

An additional analysis of the confidence values was prompted by Kulhavy's observation "that one of the reasons why corrective aspects of feedback have received so little attention is simply that many studies fail to analyze error and correct responses separately". (Kulhavy, 1976, p.221) As a result of this statement, confidence values were sorted according to whether the student's response was right or wrong. The mean confidence values for correct answers on the test and retest is presented in Table 18.

Table 18
Mean Confidence Values for Correct Answers

Group	n	Test	n	Retest
ICA	73	5.98	92	6.32
ICU	71	5.31	80	6.02
DCA	83	5.66	102	6.04
DCU	56	6.30	86	5.93

n(max) = 115

A three-way analysis of variance, with Factor C (test,retest) repeated, indicated a significant interaction between Factor A (Immediate, Delay), Factor B(CA,CU) and Factor C (test,retest) ($p. < .05$). Table 19 summarizes this finding. Subsequent post hoc analyses did not identify significant differences between the treatment groups.

Table 19
Analysis of Mean Confidence Values
for Correct Answers

SUMMARY OF ANALYSIS OF VARIANCE SOURCE	SS	DF	MS	F	P
BET SUBJ	17.535	19			
A (I,D)	0.0601	1	0.0601	0.06	0.809
B (CA,CU)	0.115	1	0.115	0.12	0.738
AB	1.409	1	1.409	1.41	0.252
SUBJ W GROUP	15.950	16	0.997		
WITHIN SUBJ	4.885	20			
C (T,RE)	0.699	1	0.699	4.25	0.056
AC	0.689	1	0.689	4.18	0.058
BC	0.087	1	0.087	0.53	0.478
ABC	0.776	1	0.776	4.72	0.045
C X SUBJ W G	2.634	16	0.165		

The mean confidence values for wrong answers on the test and retest are summarized in Table 20.

Table 20
Mean Confidence Values for Wrong Answers

Group	n	Test	n	Retest
ICA	42	5.57	23	5.70
ICU	44	4.42	35	5.30
DCA	32	4.79	13	3.29
DCU	59	5.52	29	5.55

n(max) = 115/group

A three-way analysis of variance, with Factor C (test,retest) repeated, indicated a significant interaction between Factor A (Immediate, Delay) and Factor B (CA,CU) ($p < .05$) (Table 21). Subsequent post hoc analyses ,using Scheffe's method, did not identify any signficant differences.

Table 21
Analysis of Mean Confidence Values for Wrong Answers

SUMMARY OF ANALYSIS OF VARIANCE SOURCE	SS	DF	MS	F	P
BET SUBJ	82.132	19			
A (I,D)	2.093	1	2.093	0.73	0.405
B (CA,CU)	1.307	1	1.307	0.46	0.509
AB	12.894	1	12.894	4.50	0.050
SUBJ W GROUP	45.839	16	2.865		
WITHIN SUBJ	34.618	20			
C (T,RET)	0.141	1	0.141	0.08	0.776
AC	3.875	1	3.875	2.30	0.149
BC	3.254	1	3.254	1.93	0.18
ABC	0.387	1	0.387	0.23	0.638
C X SUBJ W G	26.962	16	1.685		

Discussion: Confidence Values

In the review of literature, Sturges(1976) indicated that subjects receiving 24 hour delay of feedback would be more confident about their retention test scores than those subjects who received immediate feedback. In this study, the greatest change occurred for the immediate feedback group. The confidence of the immediate feedback group grew significantly ($p < .02$) between the end-of-chapter test and the retention test. No significant change occurred for the delay of feedback group ($p < .38$). This finding would appear to partially parallel the test score findings. Just as test scores increased between the end-of-chapter test and the retention test, an increase in confidence could be expected between these tests. The members of the delay of feedback group maintained the same relative degree of confidence in their responses on both tests. The immediate feedback group increased in mean confidence. An examination of individual confidence responses indicated some subjects were extremely confident on the end-of-chapter test --even when their response was wrong. On the retest, the subjects had higher scores and the confidence values appeared to be more in keeping with their overall increased success. It will be recalled that both the immediate and the delay feedback groups scored significantly higher on the retention test than on the end-of-chapter test ($p < .0001$). From the confidence findings, it would appear that immediate feedback resulted in a much higher confidence in answers than did

delay of feedback. This finding does not support Sturges(1976).

An examination of the confidence values for correct and wrong answers did not provide any additional information, in spite of the suggestion by Kulhavy (1977) that the separation of data into correct and wrong responses would provide more insight into the corrective power of feedback messages.

If a legitimate concern of instructional designers is to ensure that students respond with the highest possible confidence to test items, then immediate feedback would appear to be the the design construct of choice. This finding does not confirm Sturges (1976) and therefore may be an indication that this sample was more concerned about the feedback in terms of understanding the subject matter than superficially reading it to determine if their answer was correct. It is also possible that the sample, composed of graduate students, were motivated because of the future credit attached to the retest scores and therefore attended carefully to the information provided. It might also be noted that this test was several hours in length and that the students were familiar with the CAI mode of testing. Students were not rushed, nor was there any pressure for them to advance quickly.

If these findings are considered in the context of the theory of memory advanced in Chapter II, subjects in the immediate feedback group accomplished tuning and

restructuring of their schema as effectively as did the delay feedback group. In addition, the retention test confidence means indicate that the change in the confidence of the immediate feedback group may be attributable to more exacting retrieval terms and a resultant increase in the level of verification at each stage of the recursive retrieval process (Williams, 1977).

Feedback Latency

A second and additional question for investigation was to determine if the time subjects took to examine feedback (feedback latency) differed as a result of feedback timing (immediate,delay) or feedback messages (CA,CU). The feedback latency means are summarized in Table 22.

Table 22
Feedback Latency by Treatment Group

Group	Feedback Latency* Test	Retest
ICA	302.05	27.09
ICU	335.59	93.20
DCA	376.17	43.95
DCU	319.83	133.83
Mean	333.41	74.52

*Time in seconds

A three-way analysis of variance, with Factor C (test,retest) repeated, indicated a significant difference on Factor C (Table 23). Post hoc analyses indicated all treatment groups significantly reduced the mean time spent reading the feedback messages following the retention test ($p < .0001$).

Table 23
Summary of Analysis on Feedback Latency

Sources of Variation	df	MS	F
Between Subjects	39		
A (Immediate, Delay)	1	479.18	.09
B (Correct Answer, Cue)	1	2433.62	.46
AB	1	4476.00	.84
Subjects within groups	36	5298.49	
Within Subjects	40		
C (Test, Retest)	1	685837.63	130.31**
AC	1	119.81	.02
BC	1	3379.37	.64
ABC	1	2357.63	.45
Time X subjects within groups	36	5262.95	

*F.95 (1, 36) 4.11
**F.99 (1, 36) 7.39

Kulhavy (1976) suggested that high confidence correct answers would be read for direction not instruction, whereas high confidence wrong answers would be read carefully, possibly argumentatively, but would result in a low repetition rate. The Norman-Rumelhart theory of memory (discussed earlier) supports such an argument, since highly confident, but wrong responses would seem to necessitate a tuning or restructuring of the responsible schema.

In order to test Kulhavy's theory, high and low confidence responses were selected. High confidence was defined as a '7' (absolutely certain) and low confidence was defined as a range between 1-4 (not certain). These definitions each provided approximately 20% of the total responses. The high and low confidence responses were then classified as correct or wrong according to the matching

test answer. During this process two subjects were discovered who responded in a highly confident manner to all answers. These subjects may not have understood the concept of confidence or possibly did not wish to cooperate with the study. As a result, responses of these subjects were dropped from the analysis. The summary of all high confidence responses with matching feedback latencies is presented in Table 24.

Table 24
High Confidence Items and Mean Feedback Latencies*

Group		Feedback Latency Test	Retest	n Test	Retest
ICA					
CA	X	18.1	6.3	17	19
	SD	14.1	7.8		
WA	X	49.8	14.1	4	2
	SD	174.1	1.7		
ICU					
CA	X	56.2	14.6	23	25
	SD	47.7	17.5		
WA	X	127.4	70.1	7	5
	SD	88.9	61.2		
DCA					
CA	X	18.6	5.4	33	35
	SD	20.7	7.5		
WA	X	39.5	71.0	4	2
	SD	41.4	39.7		
DCU					
CA	X	32.9	2.1	13	14
	SD	42.7	1.7		
WA	X	51.2	4.8	2	1
	SD	52.8	0		

*Latency times are in seconds
n(max)=115=(23 items/test X 5 subjects/group).
ICA & DCU groups have 4 subjects each.

With reference to Table 24, it will be noted that each treatment group is divided into correct (CA) and wrong (WA) answers. The number of responses which are represented by the mean are small when it is recalled that the total number of responses per group is 115. As indicated earlier, two subjects were dropped due to indiscriminant high confidence. As a result the ICA and DCU groups have only 4 subjects each. A further examination of the data revealed that it was not uncommon to find one subject contributing most of the high confidence items for a group.

Due to the small sample size and the subjective nature of high confidence, the data were not statistically analyzed. A general examination of the data suggests that those high confidence responses which were wrong resulted in slightly longer feedback latencies. It should be noted that cue feedback, is by its nature, more verbose than correct answer feedback. The mean number of words per cue message is 30, hence a period of time is required to read a cue, whereas the correct answer feedback is already familiar. The latency times described here represent the total time the feedback message remained on the computer terminal screen. No measure is available to determine if the subjects attended continually to the message during this time period.

If feedback is to be useful it must evoke a change in the behavior responsible for wrong answers and reinforce those behaviors which are responsible for low confidence correct answers. Table 25 summarizes the low confidence

data. No statistical procedures were performed on these data due to the nature of the data.

Table 25
Low Confidence Items and Mean Feedback Latencies*

Group		Feedback Latency Test	Retest	n Test	Retest
ICA					
CA	X	11.8	15.8	6	11
	SD	8.2	21.5		
WA	X	46.0	36.3	12	7
	SD	77.5	27.1		
ICU					
CA	X	94.8	3.8	13	19
	SD	147.2	3.3		
WA	X	81.5	14.8	25	19
	SD	80.1	22.5		
DCA					
CA	X	48.9	6.2	11	19
	SD	80.6	7.5		
WA	X	45.7	10.8	13	5
	SD	54.8	7.8		
DCU					
CA	X	-	3.9	0	3
	SD	-	16.1		
WA	X	90.3	35.9	6	3
	SD	35.9	59.6		

*Latency times are in seconds
n(max)=115=(23 items/test X 5 subjects/group).
ICA & DCU groups have 4 subjects each.

Generally, the data indicate subjects read the feedback messages on the end-of-chapter test. The amount of time spent reading the feedback message following a correct or incorrect response is approximately the same. (Compensating for the fact that latencies are in seconds and the number of items is small.) Following the retention test, subjects appeared to spend slightly more time reading feedback to wrong answers than reading feedback to the correct answers.

Discussion: Feedback Latency

This study found subjects spent significantly less time reading feedback messages following the retention test than during or following the end-of-chapter test. The reasons for the decreased time to read feedback messages on the retention test may be due to a combination of (1) a desire to see the test score (which followed feedback), (2) a need to move on to the next chapter or section of the course, (3) a decision that feedback was unlikely to provide assistance in later tests, and (4) the material was, by this time, well known. It will be recalled that feedback on the retention test was provided (1) only after the last test item was answered, (2) serially in the order the items were answered, and (3) the feedback message was the question stem with correct answer underlined.

It was hoped that confidence measures, in combination with feedback latencies, would provide additional evidence for the selection of one type of feedback timing and/or one type of feedback message design. However, the data do not provide any evidence in support of any particular combination of feedback timing or feedback messages used in this study.

Response Latency

The third additional research question asked if feedback timing and feedback message design had any effect upon response latency... the time subjects required to answer a test question. The mean response times are

presented in Table 26.

Table 26
Response Latency by Treatment Group

Group	Response Latency* Test	Retest
ICA	225.09	13.61
ICU	175.25	11.48
DCA	191.94	7.06
DCU	193.72	13.13
Mean	196.50	11.25

*Time in seconds

A three-way analysis of variance, with Factor C (test,retest) repeated, was performed. The analysis indicated a significant drop in response times between the test and retest (Factor C) ($p < .01$). In addition, a significant interaction was detected between Factor B (CA,CU) and Factor C (test,retest) ($p < .05$). These results are presented in Table 27.

Table 27
Summary of Response Latency Analysis

Sources of Variation	df	MS	F
Between Subjects	39		
A (Immediate, Delay)	1	16772.0	1.37
B (Correct Answer, Cue)	1	22177.0	1.81
AB	1	5462.0	.45
Subjects within groups	36	12238.3	
Within Subjects	40		
C (Test, Retest)	1	1340475.0	164.26**
AC	1	-0-	1.00
BC	1	39960.0	4.90*
ABC	1	16161.0	1.98
Time X subjects within groups	36	8160.53	

*F.95 (1, 36) 4.11
**F.99 (1, 36) 7.39

Post hoc analyses indicated (1) all treatment groups significantly reduced the mean time required to answer the test items ($p < .001$) and (2) the cue group required a significantly higher mean time to respond to the retention test than did the immediate feedback group ($p < .01$).

Discussion: Response Latency

The increased test scores on the retention test, as well as the reduction in response times, all point to greater familiarity with the test material. Of particular note was the significant difference between subjects who received CA feedback and those who received CU feedback. It will be recalled that a significant difference existed between the mean scores of the CA and CU feedback groups on the end-of-chapter test ($p < .005$) and that this difference was not significant on the retention test ($p < .11$). The legacy of this difference appears to have carried over to the cue group on the retention test in the form of longer response times.

These findings indicate that cue feedback resulted in both an increase in mean test scores and, to some extent, increased mean processing times on the retest. Although the increased mean scores were desirable, the increased mean response times may indicate either a temporary difficulty in retrieving information for a solution (confusion from cue feedback) or (2) a more systematic verification process based upon the knowledge that this material was not well known in the past. Either of these interpretations may be

valid. The researcher favours the view that increased retention test mean response latency is an artifact of former subject matter difficulty, even though confidence values for this group did not indicate a perceived lack of certainly in the answers when compared with the CA feedback group.

V. Conclusions and Recommendations

A. Conclusions

This study was designed to investigate the common constructs used in CAI testing sessions. At present, emphasis is placed on providing feedback as rapidly as possible. CAI computer hardware configurations often emphasize the need to produce responses within 1-2 seconds. Naturally, it has been believed that feedback should also be presented with similar swiftness. Feedback messages have tended to be brief and to the point, e.g. "You are right", or "You are wrong, the answer is...". Several CAI systems have been designed to provide short feedback messages with very little programming effort. Since no evidence existed to suggest other CAI methods would be better, or that the current practices were the best, CAI authors and programmers have tended to produce courseware that was compact, efficient and required the least effort.

This study examined the effect of immediate feedback and 24 hour delay of feedback upon long term retention. On a retention test one week following an end-of-chapter test, no significant difference was found between the immediate and delay feedback groups.

The study also examined the effects of two types of feedback messages upon long term retention. One message provided a re-display of the multiple choice test item with the correct answer underlined (CA), the other feedback

message was written to provide a cue to the correct answer (CU). Briefly, this cue was a CRT display of several sentences or a formula which the statistical instructors agreed would 'cue' the subject to the correct answer. These cues were, on average, 30 words in length.

This study departed from all earlier research activity, since no other study has involved: (1) integrating the testing techniques within a stable CAI, (2) maintaining the freedom of instructors to interact with students within a CAI environment, (3) allowing students to respond to test items in their own time, (4) administering the test and retest under the same CAI conditions, (5) standardizing the interval between test and retest at 7 days for all students, (6) collecting data on the confidence students had in their answers as well as collecting data on the time taken to answer items and read feedback messages, (7) assigning course credit for both the test and retest, and (8) examining recall as well as recognition of test material.

The CA and CU feedback groups had different mean scores on the end-of-chapter test ($p < .005$). There was no difference between CA and CU feedback groups on the mean scores on the retention test one week later ($p < .11$). Also, all groups scored better on the retention test in comparison to the end-of-chapter test ($p < .0001$).

Additional questions of interest to CAI instructional designers were asked. These were with regard to the effect of feedback timing and feedback message design may have upon:

a. the mean confidence assigned by students to their responses,

b. the mean latency time require by students to produce responses and

c. the mean latency time taken to read a feedback message.

One-half of the subjects in the sample (N=20) were asked to indicate the confidence they had in their answers. The subjects responded by ranking the 'certainty of correctness' of their response to a test item on a continuum from 1 (not certain) to 7 (absolutely certain). It was found that no difference in mean confidence values existed between the four treatment groups (immediate CA, immediate CU, delay CA, delay CU) on the end-of-chapter test, or on the retention test. The immediate feedback group (CA,CU) did increase their mean confidence value between the test and retest however ($p < .02$).

An examination of the time subjects spent reading the feedback messages (feedback latency) indicated no differences existed between the treatment groups on the mean feedback latency times for the test and retest. A significant decline in the mean feedback latency times occurred for all treatment groups between the test and retest ($p < .001$).

The time subjects required to answer test items (response latency) was analyzed. A significant decline in mean response latency times occurred between the test and

retest for all groups ($p < .001$). In addition, on the retention test, the feedback message group required more time to respond on the retention test than did the CA feedback message group ($p < .01$). It was suggested the differences may be traced to the lower test scores achieved by the CU group on the end-of-chapter test.

In summary, this study, based upon a sample consisting of university graduate students, supports the current practices of providing immediate feedback using short feedback messages (indicating just the correct answer). Providing a delay in feedback was a cumbersome, time consuming process for the IBM 1500 author. This feedback mode also required more physical space in the computer program and a vigilance on the part of the instructor to ensure subjects returned within 24 hours. Subjects also seemed to prefer immediate feedback as they wished to know which of their answers were correct after completing the test. End-of-test feedback is possible, but perhaps future CAI systems will allow the subject to select the feedback mode of preference. Perhaps it should also be stated that delaying feedback did not degrade subjects' performances. However, a delay in feedback may have advantages for certain other types of learners.

The use of cues as feedback messages provided some difficulties for the CAI instructors, the programmer, and the STAT 1 students. First, the instructors did not entirely agree on the messages provided, since each has his own

approach to assisting the student. Second, cues required more programmer effort, and more space in the computer computer program. Third, some subjects did not like the cue feedback and were quite argumentative about their usefulness Finally, there was no evidence to indicate that subjects spent more time reading cues (in comparison to correct answer feedback), or that test scores increased as a result of cues. In contrast to these findings, a body of research and theory exists to support the use of cues (Sturges, 1972; Norman & Rumelhart, 1976). Further research, using different age groups, different course content, and without the high levels of instructor support or learner motivation may confirm the value of cues as feedback.

B. Recommendations

A major difficulty encountered in this study was the size of the sample. This was particularly troublesome when investigating the relationship of confidence values to feedback time and feedback message design. The reason only one-half of the available subjects were required to indicate their confidence values was because no previous research data were available to indicate the effect this procedure might have on the test scores, feedback latency, or response latency. This study found that requiring subjects to submit confidence values had no significant impact upon other variables under investigation. Had a larger sample been available, this study would have adopted the fine-grained answer analysis suggested by Kulhavy (1972) and Phye (1977).

The investigation of error perseveration and long term retention appears to be one with excellent potential. It is recommended that future studies dispense with the control group (non confidence measuring) and devote the entire sample to an in-depth study of learner behavior on test-retest paradigms.

It is also recognized that the population used for the study consists of graduate students. For the most part these students are a highly motivated, competitive group. In this case they were enrolled in a core curricular requirement. Their study habits were expected to challenge the theories and earlier research cited in the review of literature. The CAI learning environment and readily available instructor assistance may have also reduced the impact of the delay retention effect and cue type feedback messages reported by Sturges, Kulhavy and others. The findings of this study suggest continued reseach using a variety of student populations, CAI/CMI delivery modes, and differing degrees of instructor support.

Some studies have indicated cue feedback enhances long term retention. Since this study did not confirm these earlier findings, it would appear more needs to be known about the construction of cues if they are to have the effectiveness claimed by some researchers. Perhaps a variety of cues could be provided for each test item and the student tracked by the computer to determine which ones are found most useful. From the feedback latency data it seems that

students examine in detail the messages provided, irrespective of the feedback mode (immediate or delayed). More needs to be known about creating conditions of learning within the evaluation process. It is recommended that CAI designers experiment with variable length feedback messages following incorrect responses to determine the effectiveness of learning sessions imbedded in evaluation.

Finally, it is recommended that a study be carried out to determine the impact of providing a retention test which states previous test results as part of the feedback message. That is, in addition to feedback related to correcting the error, the subject would be told his response history on the item, e.g. "You got this question wrong last time and the answer you used was 'x'" or "You got this question right the last time, but this time you are wrong...the answer is 'x'". A second retention test would examine the impact of stating the learner's history of responses to determine if this triggers longer feedback latency and a positive change in behavior (correct answer).

References

Abra, J.C. List-1 unlearning and recovery as a function of the point of interpolated learning. Journal of Verbal Learning and Behavior, 1969, 8, 494-500.

Adams, J.A. Response feedback and learning. Psychological Bulletin. 1968, 70, 486-504.

Ammons, R.B. Effects of knowledge of performance: a survey and tentative theoretical formulation. Journal of General Psychology. 1956, 54, 279-299.

Anderson, R.C., & Faust, G.W. The effects of strong formal prompts in programmed instruction. American Educational Research Journal. 1967, 4, 345-352.

Anderson, R.C., Kulhavy, R.W., & Andre, T. Feedback procedures in programmed instruction. Journal of Educational Research, 1971, 62, 148-156.

Anderson, R.C., Kulhavy, R.W., & Andre, T. Conditions under which feedback facilitates learning from a programmed lesson. Journal of Educational Psychology, 1972, 63, 186-188.

Annet, J. The role of knowledge of results in learning: A survey. In J.P. DeCecco (Ed.), Educational Technology. New York: Holt, Rinehart & Winston, 1964.

Angell, G.W. The effect of immediate knowledge of quiz results on final examination scores in freshman chemistry. Journal of Educational Research, 1949, 42, 391-394.

Bartlett, F. C. Remembering. Cambridge: Cambridge University

Press; 1932.

Bilodeau, I.M. Information feedback. In E.A. Bilodeau (Ed.), Acquisition of skill. New York: Academic Press, 1966.

Bourne, L.E. Effects of delay of information feedback and task complexity on identification of concepts. Journal of Educational Psychology, 1957, 54, 201-207.

Brackbill, Y. The impairment of learning under immediate reinforcement. Journal of Experimental Child Psychology, 1964, 1, 199-207.

Brackbill, Y., Bravos, A., & Starr, R.H. Delay improved retention of a difficult task. Journal of Comparative and Physiological Psychology, 1962, 55, 947-952.

Brackbill, Y. & Kappy, M.S. Delay of reinforcement and retention. Journal of Comparative and Physiological Psychology, 1962, 55, 14-18.

Brackbill, Y., Wagner, J., & Wilson, D. Feedback delay and the teaching machine. Psychology in the Schools, 1964, 1, 148-156.

Brown, A.L. The development of memory: Knowing about knowing and knowing how to know. In H. Reese (Ed.), Advances in child development and behavior, Vol. 10: New York: Academic Press, 1975.

Cermak, L.S. Human memory research and theory. New York: Ronald Press, 1972.

Craik, F.I.M., & Tulving, E. Depth of processing and the retention of words in episodic memory. Journal of Experimental Psychology: General, 1975, 104, 268-294.

Elley, W.B. The role of errors in learning with feedback. British Journal of Educational Psychology, 1966, 36, 296-300.

English, R.A., & Kinger, J.R. The effect of immediate and delayed feedback on the retention of subject matter. Psychology in the Schools, 1966, 3, 143-147.

Fitts, P.M. Perceptual-motor skill learning. In A.W. Melton (Ed.), Categories of human learning. New York: Academic Press, 1964.

Gilman, D.A. Comparison of several feedback methods for correcting errors by computer-assisted instruction. Journal of Educational Psychology, 1969, 60, 503-508.

Glaser, R. & Cooley W.W. Instrumentation for teaching and instructional management. In R.M.N. Travers (Ed.) Second handbook of research on teaching Chicago: Rand McNally, 1973, 832-857.

Gray, R.T. An evaluation of the effect of an immediate feedback device used with typical college classroom tests, 1968, (ERIC, BR-6-8156-015-658).

Hansen, J.B. Effects of feedback, learner control, and cognitive abilities on state anxiety and performance in a computer-assisted instruction task. Journal of Educational Psychology, 1974, 66, 247-254.

Hewitt, C., Bishop, P., & Steiger, R. A universal modular ACTOR formalism for artificial intelligence. In Proceedings of the third international conference on artificial intelligence, Stanford: 1977.

Holland, J.G., & Skinner, B.F. The analysis of behavior. New York: McGraw-Hill, 1961.

Hunka, S., Romaniuk, G. & Maguire, T. Report on the use of the computer-assisted instruction course STAT1 as used in Educational Psychology 502, 1975-1976. Edmonton: Division of Educational Research Services, University of Alberta, 1976.

Judd, C.A. Practice without knowledge of results. Psychological Review Monograph Supplement. 1905-6, 7, 185-198.

Kaess, W., & Zeaman, D. Positive and negative knowledge of results in a Pressey-type punchboard. Journal of Experimental Psychology, 1960, 60, 12-17.

Kaufman, R.A. Experimental evaluation of the role of remedial feedback in an intrinsic program. Journal of Programmed Instruction, 1963, 2, 21-30.

Kearsley, G. Some facts about CAI: Trends 1970-1976. Edmonton: Division of Educational Research Services, University of Alberta, 1976.

Keppel, R.W. Retroactive and proactive inhibition. In T.R. Dixon & D.L. Herton (Eds.), Verbal behavior and general behavior theory. Engelwood Cliffs, N.J.: Prentice Hall, 1968.

Krumboltz, J.E., & Weisman, R.G. The effect of intermittent confirmation in programmed instruction. Journal of Educational Psychology, 1962, 53, 250-253.

Kulhavy, R.W. Feedback in written instruction. Review of Educational Research, 1977, 47, 211-232.

Kulhavy, R.W., & Anderson, R.C. Delay-retention effect with multiple-choice tests. Journal of Educational Psychology, 1972, 63, 505-512.

Kulhavy, R.W., & Parsons, J.A. Learning-criterion error perseveration in text materials. Journal of Educational Psychology, 1972, 63, 81-86.

Kulhavy, R.W., & Swenson, I. Imagery instructions and the comprehension of text. British Journal of Educational Psychology, 1975, 45, 47-51.

Kulhavy, R.W., Yekovich, F.R., & Dyer, J.W. Feedback and response confidence. Journal of Educational Psychology, 1976, 68, 522-528.

Kulhavy, R.W., & Yekovich, F.R. Feedback in instruction. Chapter in Encyclopedia of Instructional Development. San Diego: Navy Personnel Research and Development Center, in press.

Leherissey, B.L., O'Neil, H.F., Heinrich, D.L., & Hansen, D.N. Effect of anxiety, response mode, subject matter familiarity, and program length on achievement in computer-assisted learning. Journal of Educational Psychology, 1973, 64, 310-324.

Leherissey, B.L., O'Neil, H.F., & Hansen, D.N. Effects of memory support on state anxiety and performance in computer-assisted learning. Journal of Educational Psychology, 1971, 62, 413-420.

Linton, M. Memory for real world events. In D.A. Norman and D.E. Rumelhart Explorations in Cognition. San Francisco:

Freeman, 1975, 376-404.

Lovayne, H., & Lucas, J. The memory book. New York: Baltimore Books, 1974.

Lublin, S.C. Reinforcement schedules, scholastic aptitude, autonomy need and achievement in a programmed course. Journal of Educational Psychology, 1965, 56, 295-302.

Markle, S.M. Good frames and bad. New York: Wiley, 1964.

Markowitz, N., & Renner, K.E. Feedback and the delay-retention effect. Journal of Experimental Psychology, 1966, 72, 452-455.

Mayer, R.E. Information processing variables in learning to solve problems. Review of Educational Research, 1975, 45, 525-542.

McDonald, F.J., & Allen, D. An investigation of presentation, response and correction factors in programmed instruction. Journal of Educational Research, 1962, 55, 502-507.

Melching, W.H. Programmed instruction under a feedback schedule. National Society for Programmed Instruction Journal, 1966, 5, 14-15.

Merrill, M.D. Correction and review on successive parts in a learning hierarchical task. Journal of Edycational Psychology, 1965, 65, 225-234.

Moore, J.W., & Smith, W.I. Knowledge of results in self-teaching spelling. Psychological Reports, 1961, 9, 717-726.

Moore, J.W., & Smith, W.I. Role of knowledge of results in

programmed instruction. Psychological Reports, 1964, 14, 407-423.

More, A.J. Delay of feedback and the acquisition and retention of verbal materials in the classroom. Journal of Educational Psychology, 1969, 60, 339-342.

Newell, A. Production systems: Models of control structures. In W. G. Chase (Ed.), Visual information processing. New York: Academic Press, 1973.

Newman, M.I., Williams, R.G., & Hiller, J.H. Delay of information feedback in an applied setting: Effects on initially learned and unlearned items. Journal of Experimental Education, 1974, 42, 55-59.

Norman, D.A., & Bobrow, D.G. On the role of active memory processes in perception and cognition. In D.N. Cofer (Ed.), The structure of human memory. San Francisco: Freeman, 1976.

Norman, D. A., & Bobrow, D.G. Descriptions: A basis for memory acquisition and retrieval(CHIP 74). San Diego: University of California, Center for Human Information Processing, 1977.

Norman, D.A., & Rumelhart, D.E. Accretion, tuning and restructuring: three modes of learning. San Diego: University of California, Center for Human Information Processing. 1976.

O'Neil, H.F., Jr. Effects of stress on state anxiety and perform computer-assisted learning. Journal of Educational Psychology, 1972, 65, 473-481.

Paige, D.D. Learning while testing. Journal of Educational Research, 1966, 59, 276-277.

Phye, G.D. The role of informative feedback in productive learning, New York, N.Y.: 1977. (ERIC ED 138612).

Phye, G., & Baller, W. Verbal retention as a function of the informativeness and delay of informative feedback: A replication. Journal of Educational Psychology, 1970, 61, 380-381.

Phye, G., Eugliemella, J., Sola, J. Effects of delayed retention on multiple -choice test performance. Continuing Educational Psychology, 1976, 1, 26-36.

Reddy, R. & Newell, A. Knowledge and its representation in a speech understanding system. In L.W. Gregg (Ed.), Knowledge and cognition. Hillsdale, N.J.: Laurence Erlbaum Associates, 1974.

Renner, K.E. Delay of reinforcement: A historical review. Psychological Bulletin, 1964, 61, 341-361.

Roe, A.A. A comparison of branching methods for programmed learning instruction. Journal of Educational Research, 1962, 55, 407-416.

Rosenstock, E.H., Moore, W.J., & Smith, W.I. Effects of several schedules of knowledge of results on mathematics achievement. Psychological Reports, 1965, 17, 535-541.

Sassenrath, J.M. Effects of delay of feedback on retention of prose material. Psychology in the Schools, 1972, 9, 194-197.

Sassenrath, J.M. Theory and results on feedback and

retention. Journal of Educational Psychology, 1975, 67, 894-899.

Sassenrath, J.M., & Garverick, C.M. Effects of differential feedback from examinations on retention and transfer. Journal of Educational Psychology, 1965, 56, 259-263.

Sassenrath, J.M., & Yonge, G.D. Delayed information feedback, feedback cues, retention set, and delayed retention. Journal of Educational Psychology, 1968, 59, 69-73.

Sassenrath, J.M., & Yonge, G.D. Effects of delayed information feedback and feedback cues in learning and retention. Journal of Educational Psychology, 1969, 60, 174-177.

Sax, G. Concept acquisition as a function of differing schedules and delays of reinforcement. Journal of Educational Psychology, 1960, 51, 32-36.

Selfridge, O.G., & Neisser, M. Pattern recognition by machine. Scientific American, 1960, 203, 60-68.

Skinner, B.F. The technology of teaching. New York: Appleton-crofts, 1968.

Strong, H.R., & Rust, J.O. The effects of immediate knowledge of results and task definition on multiple-choice answering. Journal of Experimental Education, 1973, 42, 77-80.

Sturges, P.T. Verbal retention as a function of the informativeness and delay of information feedback. Journal of Educational Psychology, 1969, 60, 11-14.

Sturges, P.T. Information delay and retention: Effect of information on feedback and tests. Journal of Educational Psychology, 1972, 63, 32-43.(a)

Sturges, P.T. Effect of instructions and form of informative feedback on retention of meaningful material. Journal of Educational Psychology, 1972, 63, 99-102.(b)

Sturges, P.T. Delay of informative feedback and computer managed instruction. (Semi-annual technical report) San Diego: Naval Personnel, Research and Development Center, 1976.

Sturges, P.T. Immediate vs. delayed feedback in a computer managed test: effects on long term retention. San Diego: Naval Personnel, Research and Development Center, 1978.

Sturges, P.T., Sarafino, E.P., & Donaldson, P.I. The delay-retention effect and informative feedback. Journal of Educational Psychology, 1968, 78, 357-358.

Sullivan, H.J., Baker, R.L., & Schutz, R.E. Effect of intrinsic and extrinsic reinforcement contingencies on learner performance. Journal of Educational Psychology, 1967, 58, 165-169.

Sullivan, H.J., Schutz, R.E., & Baker, R.L. Effect of systematic variations in reinforcement contingencies on learner performance. American Educational Research Journal, 1971, 8, 135-142.

Surber, J.R., & Anderson, R.C. Delay-retention effect in natural classroom settings. Journal of Educational Psychology, 1975, 67, 170-173.

Tait, K., Hartley, J.R., & Anderson, R.C. Feedback

procedures in computer assisted arithmetic instruction. British Journal of Educational Psychology, 1973, 43, 161-171.

Talyzina, N.F. Psychological bases of programmed instruction. Instructional Science, 1973, 2, 243-280.

Travers, R.M.W., Van Wagenen, R.K., Haygood, D.H., & McCormick, M. Learning as a consequence of the learner's task involvement under different conditions of feedback. Journal of Educational Psychology, 1964, 55, 167-173.

Tulving, E. Episodic and semantic memory. In E. Tulving, & W. Donaldson (Eds.) Organization of memory. New York: Academic Press, 1972.

Underwood, B.J., & Ekstrand, B.R. Studies of distribution of proactive inhibition. Journal of Educational Psychology, 1967, 74, 574-580.

Wentling, T.L. Mastery versus nonmastery instruction with varying test item feedback treatments. Journal of Educational Psychology, 1973, 65, 50-58.

Williams, M.D. The process of retrieval from very long term memory. Unpublished doctoral dissertation, University of California, San Diego, 1977.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

APPENDIX A

1. An example of one of the test items.

A t-test can be used to:

a. compare two means
b. correlate two variables
c. draw two random samples
d. compare two independent variances

Type the letter of your answer

2. Confidence measures are presented to those required to respond. In this example, if answer option 'a' is selected then the confidence questions would be the following four.

1. How confident are you that your answer is correct?

a. compare two means

```
Not                                          Absolutely
Certain                                        Certain
I------I------I------I------I------I------I------I
1      2      3      4      5      6      7
```

Point to the number on the rating scale

1. How confident are you that your choice is wrong?

b. correlate two variables

```
Not                                          Absolutely
Certain                                        Certain
I------I------I------I------I------I------I------I
1      2      3      4      5      6      7
```

Point to the number on the rating scale

1. How confident are you that your choice is wrong?

c. draw two random samples

```
Not                                          Absolutely
Certain                                        Certain
I------I------I------I------I------I------I------I
1      2      3      4      5      6      7
```

Point to the number on the rating scale

1. How confident are you that your choice is wrong?

d. compare two independent variances

```
Not                                          Absolutely
Certain                                        Certain
I------I------I------I------I------I------I------I
1      2      3      4      5      6      7
```

Point to the number on the rating scale

3. Irrespective of whether confidence measures are required, the subject receives feedback in one of two forms. The first example is of the correct answer underlined, the second, a cue. The delivery of these messages is either immediately after the last confidence measure or after answering the question (if no confidence measures are taken), or 24 hours later when the subject signs on to the IBM 1500 system.

Feedback message: correct answer underline

A t-test can be used to:

a. compare two means
b. correlate two variables
c. draw two random samples
d. compare two independent variances

The answer is underlined above

Feedback message: cue

First the question is presented again.

A t-test can be used to:

a. compare two means
b. correlate two variables
c. draw two random samples
d. compare two independent variances

Second the cue is presented.

The following statements may be helpful:

1. A t-test may be used with independent or dependent groups.
2. A t-test of two independent groups is the equivalent of an ANOVA for two independent groups.
3. The numerator is $(\bar{x}_1 - \bar{x}_2)$

XX Show me the question again

XX Move me on to the next question

** Point to one of the options above

APPENDIX B

Question 1

Item

In testing a statistical hypothesis, it is necessary to find the sampling distribution of the test-statistic. This sampling distribution is found by assuming that:

a) the hypothesis being tested is not true
b) the confidence interval -1.96 to 1.96
c) the hypothesis being tested is true
d) the sampling distribution is a normal curve

Cue

To determine the correct answer consider the following:

a) the random sampling distribution of a test-statistic must be known or assumed before analysis my proceed

b) a positive statement of purpose reflects this assumption

Question 2

Item

A t-test can be used to:

a) compare two means
b) correlate two variables
c) draw two random samples
d) compare 2 independent variances

Cue

The following statements may be helpful:

1. A t-test may be used with independent or dependent groups.
2. A t-test of two independent groups is the equivalent of an ANOVA for two independent groups
3. The numerator is $(\bar{x}_1 - \bar{x}_2)$

Question 3

Item.

In calculating the confidence interval for a population mean, one does not require:

a) standard deviation of sample scores
b) the size of the sample
c) the value of the population mean
d) the confidence level decided upon

Cue

A sample of 64 is drawn from a population whose mean is 104 and the sample standard deviation is 9. The 95% confidence interval is approximately

$$104 - 1.96 \times \frac{9}{\sqrt{64}} \leq \mu \leq 104 + 1.96 \times \frac{9}{\sqrt{64}}$$

Question 4

Item

The standard error of the mean is jsut another name for the standard deviation of:

a) a sample
b) the random sampling distribution of means
c) the random sampling distribution of any statistic
d) none of the above

Cue

1. A standard error is always a standard deviation which is decriptive of the variability of a statistic over repeated samplings.
2. The standard error in this question is specifically descriptive of what?

Question 5

Item

In the denominator of the formula for the variance of the sample, the sample size is customarily reduced by 1. The reason for this is that it makes the variance of the sample an estimate of the population variance. This variance is considered to be:

a) consistent in the critical region
b) invariant
c) an appropriate test statistic
d) unbiased

Cue

This may help:

When we estimate the population variance by $\frac{\Sigma(x-\bar{x})^2}{n}$, we fail to take into consideration that the sample mean $\bar{x}$ will be randomly different from μ. This means that our estimate is usually too small.

Question 6

Item

You have just calculated the confidence interval for the population mean with $\alpha=0.05$. You should state that:

a) insufficient evidence has been obtained to reject the level of confidence
b) differences between the sample mean and zero will exceed these limits 5% of the time
c) in 95% of such problems, the population mean will lie in such an interval
d) none of the above

Cue

Here is a statement ($.90 \leq \mu \leq .96$)
Now either the statement is true, i.e. μ is in the interval or it is false, μ is not in the interval.

Here is a confidence interval

$$p(.90 \leq \mu \leq .96) = .95$$

Question 7

Item

Although the investigator does not know it, the boys and girls in the population are equally capable of learning lesson 7. The probability that any t-test will result in the conclusion that boys are different from girls in this respect should be indicated by:

a)level of significance
b)efficiency of a statistic
c)power of the t-test
d)probability of type II error

Cue

1. The t-test is used to determine the probability of a difference in two sample means occurring by chance.
2. The probability level is that selected by the researcher; for example: 10%, 5%, or 1%.

Question 8

Item

In a controlled experiment with 12 subjects in each of two groups, a researcher used a t-test when he could have used a z-test. What was the effect of this mistake.

a) increased power
b) increased probability of rejecting a false null hypothesis
c) decreased the probability of Type I error
d) increased the probability of rejecting a true null hypothesis

Cue

1. Regardless of size, the sampling distribution of z is normal
2. With small sample sizes the t-test is distributed somewhat like z but with a curve a little fatter at the tails.
3. Type I error is the rejections of H_0 when H_0 is true

Question 9

Item

The probability distributions of the test statistic, t , with 10 degrees of freedom and 20 degrees of freedom are:

a) identical
b) identical in central tendency, but the variance of the latter is smaller
c) identical in central tendency, but the variance of the latter is greater
d) difference both in central tendency and in variance.

Cue

A population may be estimated using varying sample sizes. note the decrease in the critical value of the test statistic as the sample size increases.

Question 10

Item

A single test is administered to the same sample of 49 students on two successive Mondays. The mean of the differences in the scores is +4.0 and σ^2 (unbiased estimator) for these differences is 8.0. Assuming that the assumptions needed to satisfy the t-test are met, test the hypothesis that the population means on the two occasions are equal.

a. difference significant at .01 level
b. difference significant at .05 level but not at .01 level
c. difference not significant at .05 level
d. insufficient information is provided

Cue

You may test for significance between two means by testing whether the mean difference ($\bar{D}$) is significantly different from 0.

The sampling variance of $\bar{D}$ is found by $\frac{s^2}{N} = .163 = s^2_{\bar{D}}$ $\quad t_{obs} = \frac{\bar{D}}{s_{\bar{D}}}$

df=48, α=.05, 't'$_{.975}$=2.002

Question 11

Item

A research assistant does a pooled variance t-test, but should have done a correlated t-test. The t he obtained will be:

a) appropriate
b) smaller than it should have been
c) larger than it should have been
d) inappropriate, but nothing can be said about its size

Cue

For equal sized groups, the denominator looks like this:

1. pooled variance $\frac{S_1^2 + S_2^2}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right) = \left(\frac{S_1^2 + S_2^2}{n}\right)^{\frac{1}{2}}$

2. correlated t $\left(\frac{S_1^2 + S_2^2 - 2r_{12}S_1S_2}{n}\right)^{\frac{1}{2}}$

Remember 'r' doesn't always have to be positive, i.e. in 2

Question 12

Item

Perkins (AERA Journal, 1964) decided to use $\alpha = 0.05$. He carried out 126 independent t-tests. If the null hypothesis is true, by chance of sampling you would expect him to find approximately how many significant t's?

a) none
b) 1
c) 5
d) 6

Cue

An α of 0.05 means that we accept the probability of 5 chances in 100 of being wrong when we reject true null hypothesis. How many times would we be wrong in 126 chances.

Question 13

Item

When the pooled variance t-test is used with a sample of 15 cases and another of 25 cases, the degrees of freedom are:

a) 14
b) 38
c) 24
d) 39

Cue

Check out this formula:

$$\sqrt{\frac{\Sigma(x-\bar{x})_1^2 + \Sigma(x-\bar{x})_2^2}{n_1 + n_2 - 2}}$$

Question 14

Item

If a Pearson r = +.05 is obtained, which of the following would be proper?

a) 95% of the repeated samples would be in the range ±.05
b) the variance of one variable attributable to the other is negligible
c) high scores on one variable are paired with low scores on the other
d) although not statistically significant, the obtained r, in practice, could still be very important

Cue

r^2 is often called "variance accounted for" because $r^2 = \frac{SP^2}{SS_x SS_y}$ and $\hat{y} = bX + a$

and so y is just X in disguise

Question 15

Item

The advantage of transforming r to Fisher's z when testing the significance of an obtained r, is that the sampling:

a) variance of z is smaller than that of r
b) the shape of the z's distribution is dependent on the population value of r
c) distribution of z is approximately normal
d) distribution of z is independent of sample size

Cue

When we have a population value $\rho = .93$. The sampling distribution of r is skewed to the left because no values of r greater than 1 are possible, where as values as low as -1 are possible. When the population value $\rho = 0$, the sampling distribution r is symetric. Does the use of Fisher's z overcome this problem?

Question 16

Item

Given that in a sample of 12 observations, from a bivariate normal distribution, the sample r = +.20. Therefore, the 95% confidence interval for the population correlation coefficient is approximately:

a) (-.45, .86)
b) (-.20, .20)
c) (-.42, .69)
d) none of the above

Cue

1. If r = .20 Z_r = ? (Table E, Furguson)
2. Standard Error of $Z_r = \frac{1}{\sqrt{N-3}}$
3. 95% confidence limits = $Z_r \pm 1.96 S_{Z_r}$
4. Did you convert Z_r back to r?

Question 17

Item

As the correlation becomes smaller, the standard error of the difference between the means of the correlated variables will,

a) become larger
b) remain unchanged
c) become smaller
d) can not tell from the information given

Cue

The standard error of the difference between means for unrelated samples is:

$$\sqrt{S_1^2 + S_2^2 - 2r_{12} S_1 S_2}$$

What happens to this as r gets smaller (say for 1.00 to .5 to -.5)?
Use some plausible values for S_1 & S_2

Question 18

Item

If the number of degrees of freedom is infinite, the t distribution will be:

a) a normal distribution
b) a chi-square distribution
c) an F distribution
d) a Poisson distribution

Cue

Use your tables to see the two tailed 't' values for α = .1, .05, and .01 as N approaches ∞.

Question 19

Item

Group	Results	N
I	1, 3	2
II	7, 7, 8, 10	4

How many degrees of freedom are there for testing $H_0: u_I = u_{II}$?

a) 1
b) 4
c) 5
d) 6

Cue

For two independent groups what is the denominator in culculating the standard error of the differences between means?

Question 20

Item

Group	Results	N
I	1, 3	2
II	7, 7, 8, 10	4

In what interval does the pooled variance estimate fall?

(find $\hat{\sigma}^2$ pooled, not $\hat{\sigma}^2_{\bar{x}_I - \bar{x}_{II}}$)

a) 0.00 - 1.25
b) 1.26 - 1.75
c) 1.76 - 3.00
d) above 3.00

Cue

$$\hat{\sigma}^2 \text{ pooled} = \frac{\Sigma(x - \bar{x}_1)^2 + \Sigma(x - \bar{x}_2)^2}{n_1 + n_2 - 2}$$

Question 21

Item

Group	Results	N
I	1, 3	2
II	7, 7, 8, 10	4

Test the null hypothesis H_o: $\mu_I = \mu_{II}$

a) reject H_o at $\alpha = .10$, but not $\alpha = .05$
b) reject H_o at $\alpha = .05$, but not $\alpha = .01$
c) reject H_o at $\alpha = .01$
d) do NOT reject H_o at $\alpha = .10$

Cue

1. If $df = n_1 + n_2 - 2$
2. The standard error is:

$$\sqrt{\frac{\hat{\sigma}^2 \text{ pooled}}{n_1} + \frac{\hat{\sigma}^2 \text{ pooled}}{n_2}}$$

3. 't' = $\dfrac{\bar{x}_1 - \bar{x}_2}{\text{Standard Error}}$

Question 22

Item

A group of 9 students was tested before receiving treatment and after receiving a treatment.

$\bar{x} = 10$, $\bar{y} = 12$, $\sigma^2_x = 225$, $\sigma^2_y = 196$

$r_{xy} = .6$

What is the value of the estimate of the variance of the sampling distributions if the difference between the means.

a) 0.0 - 5
b) 5.1 - 15
c) 15.1 - 25
d) 25.1 - 200

Cue

1. Note that this is the difference between the means.

The solution looks like this . . .

$$S^2_{\bar{x}_1 - \bar{x}_2} = \frac{S_1^2 + S_1^2 - 2r_{12} S_1 S_2}{N}$$

Question 23

A group of 9 students was tested before receiving treatment and after receiving a treatment

$\bar{x} = 10$, $\bar{y} = 12$, $s^2_x = 225$, $s^2_y = 196$

$r_{xy} = .6$

In what interval does the upper critical limit of 95% confidence interval fall?

a) 1.50 - 2.00
b) 2.01 - 3.50
c) 3.51 - 4.00
d) above 4.00

Cue

1. You have part of the solution from the previous question ($S^2_{\bar{x}}$)
2. Consider this:

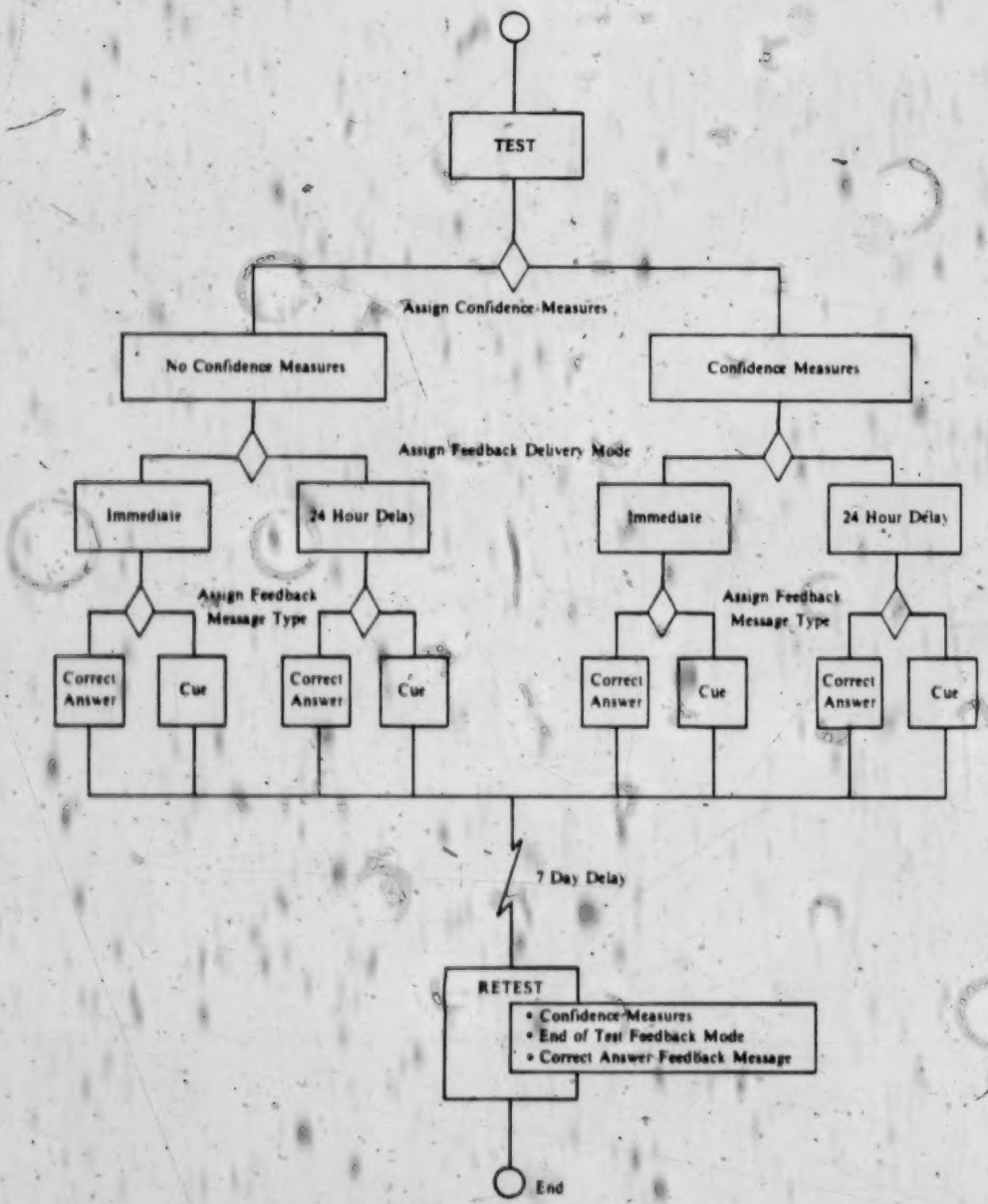
$$(\bar{x}_1 - \bar{x}_2) - t_{.975}\, S_{\bar{x}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + t_{.975}\, S_{\bar{x}}$$

3. What is the critical upper value?

On the seven day retest, the following data were used for questions 19, 20, and 21:

Group	Results	N
I	1,3,8	3
II	4,6,8,10	4

The remainder of the retest questions remained the same.

APPENDIX C

RESEARCH DESIGN